PhD-FSTM-2024-071

The Faculty of Science, Technology and Medicine

DISSERTATION

Defence held on 26/09/2024 in Esch-sur-Alzette

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG EN SCIENCES DE L'INGÉNIEUR

by

DAMIAN MINGO NDIWAGO

Born in Buea (Cameroon)

BAYESIAN MODEL SELECTION AND PRIOR IMPACT ASSESSMENT WITH A FOCUS ON DYNAMICAL SYSTEMS

Dissertation defence committee

Dr. Jack S. Hale, dissertation supervisor

Research Scientist, Université du Luxembourg

Dr. Niklas Linde

Professor, University of Lausanne

Dr. Christophe Ley, Chair

Associate Professor, Université du Luxembourg

Dr. Fatemeh Ghaderinezhad

Senior Data Scientist, TVH Parts NV

Dr. Françoise Kemp, Vice Chair

Deputy, Chambre des députés du Grand-Duché de Luxembourg

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Damian Mingo Ndiwago

Acknowledgements

I sincerely thank my supervisor, Dr. Jack S. Hale, for the research opportunity, daily supervision, and discussions about life after the PhD. I am equally grateful to Prof. Christophe Ley for his cosupervision and for organising team activities that helped me explore and appreciate Luxembourg. Special thanks to Dr. Remko Nijzink and Dr. Stanislaus Schymanski for their insightful discussions on hydrological modelling and for participating in my CET meetings. I also appreciate my visits to LIST when the pandemic slowed down.

I sincerely thank Prof. Andreas Zilian, a member of my CET, for our stimulating conversations and his role as the coordinator of the Doctoral Training Unit in Computational Modelling and Applications (DRIVEN), of which I am a proud PhD student. Prof. Andreas Zilian and Dr. Jack S. Hale were crucial in ensuring I could complete my PhD comfortably and successfully. I thank the FNR Luxembourg National Research for funding the research under the PRIDE programme (PRIDE17/12252781).

I thank the MIDAS team members for their valuable feedback and enriching conversations. I am particularly grateful to Senthil, Guendalina, Ola, Gabriel, Florian, Sophia, Katarzyna, Georgi, Gabriella, and Tianxiao for our time together and the memories we shared, especially from our conference in Berlin.

I also want to express my gratitude to Dr. Jack S. Hale research team members for their wonderful feedback and the time we shared. Thank you to Sona, Andrey, and Michal. A special thank you to Odile for all the administrative support. Thank you to the Fußballers, some of whom I shared an office with. Thanks to Saurabh, Arnaud, Chintankumar, Surendran, Natascha, Zhaoxiang, Aravind, Paris, Geremy, Thomas, Diego, Jinyuan, Roman, Parisa, and other colleagues on the fourth floor of Maison de Nombre. Lastly, I am grateful to all my friends and anyone I have crossed paths with during this journey. Special thanks to Parviel and many others who have been part of this journey. Finally, I am sincerely grateful to my parents, Ndiwago Nkwanda and Ghogomu Bernice, my brothers Cosmas, Chukang, and Victor, and my sister Victorine.

Damian Mingo Ndiwago,

Belval, September 2024

Abstract

There are many models for prediction. These models differ in the number of parameters and therefore scientists are faced with the problem of model selection. Model selection techniques seek a simple model with similar accuracy to complex models for in-sample data. While the Bayesian approach to parameter estimation is frequently used, fully Bayesian model selection is seldom used because of the high computational cost of computing the marginal likelihood, a key component of Bayesian model selection. This thesis introduces a gradient-based algorithm, Replica exchange Hamiltonian Monte Carlo (REHMC), which accurately computes the marginal likelihood when used with thermodynamic integration (TI). It also examines the often-overlooked impact of prior choices in Bayesian analysis on model outcomes, especially in Ordinary differential equation (ODE) models. The thesis extends prior impact assessment to models with more than two parameters using algorithms from computational optimal transport. It introduces a new interpretable prior impact measure based on the Wasserstein Impact Measure (WIM). Power posteriors are used to provide insights into the transitions from prior to posterior distributions. The source codes are made publicly available to encourage their adoption.

Contents

List of Figures

List of Tables

Glossary

ABC Approximate Bayesian computation. 30

AIC Akaike information criteria. 22, 47

BF Bayes factor. xiii, xiv, 10, 18, 22, 23, 24, 25, 26, 29, 31, 41, 44, 46, 47, 52, 54, 56, 58, 65, 92, 110

BIC Bayesian information criterion. 22

CI credible interval. 46

DIC deviance information criterion. xiv, xviii, 18, 43, 47, 48, 54, 55, 56, 58, 63, 65

DREAM differential evolution adaptive Metropolis. 11, 24

ESS effective sample size. 58

GLMs Generalized linear models. 15

HBV Hydrologiska Byråns Vattenbalansavdelning. 12, 13, 25, 26, 28

HMC Hamiltonian Monte Carlo. xii, 11, 24, 25, 36, 37, 38, 39, 40, 43, 65

IAT integrated autocorrelation time. xviii, 37, 45, 46, 54, 57, 61

KGE Kling Gupta efficiency. 30, 58, 65, 116

KL Kullback–Leibler. 15, 69, 70, 94

MALA Metropolis-adjusted Langevin algorithm. xii, 25, 43, 58, 65

MCMC Markov Chain Monte Carlo. 35, 36, 69, 72, 96, 97

MOPESS mean observed effective sample size. 70

NSE Nash Sutcliffe efficiency. 30, 58, 65

NUTS No-U-Turn sampler. xii, 24, 25, 42, 43, 58, 65, 76, 77, 96

ODE Ordinary differential equation. v, 12, 13, 18, 23, 26, 38, 40, 41, 69, 71, 72

ODEs Ordinary differential equations. 12, 13, 14, 15, 17, 18, 25, 27, 28, 38, 68, 69, 70, 71, 82, 89, 114, 115

OTT Optimal Transport Tools. 76, 97

PCPPP prior calibrated posterior predictive p-value. 26, 32, 52, 58, 65

pHMC preconditioned Hamiltonian Monte Carlo. xii, 25, 33, 36, 39

PMC population Monte Carlo. 36

PP probabilistic programming. 39

PPC posterior predictive check. 31

PPL probabilistic programming language. 39, 40

PPP posterior predictive p-value. 31, 32, 48, 52, 58

REHMC Replica exchange Hamiltonian Monte Carlo. v, 16, 33, 36, 37

REMC Replica exchange Monte Carlo. xii, 25, 26, 33, 36, 40, 65

REpHMC Replica exchange preconditioned Hamiltonian Monte Carlo. xii, xvii, 25, 26, 33, 36, 37, 41, 42, 43, 44, 45, 57, 64, 65

RWM random walk Metropolis. 11, 24, 36, 38, 58

SEIR Susceptible-Exposed-Infected-Removed. xix, 13, 14, 71, 75, 82, 86, 89

sWIM prior scaled Wasserstein Impact Measure. 17, 18, 68, 70, 71, 73, 74, 76, 77, 78, 80, 82, 84, 86, 88, 89, 115, 116

TFP TensorFlow probability. 25, 39, 40, 65, 96, 97

TI thermodynamic integration. v, xii, 16, 25, 26, 33, 34, 35, 36, 40, 45, 65

WAIC widely applicable information criterion. xiv, xviii, 18, 43, 47, 48, 54, 55, 56, 58, 63, 65

WIM Wasserstein Impact Measure. v, xviii, 15, 17, 68, 70, 71, 73, 76, 77, 80, 81, 82, 84, 89, 106

Chapter 1

Introduction

1.1 Background

Understanding dynamical systems is important for decision-making. A dynamical system is one whose state evolves over time (Arrowsmith & Place, 1990). These systems occur in different fields, such as ecology (Schaffer, 1985), hydrology (Machac et al., 2016), mechanics (Awrejcewicz, 2014), epidemiology (Gibson et al., 2023; Kemp et al., 2021) and atmospheric sciences (Arrowsmith & Place, 1990). Scientists use models to study these systems, but models are representations of reality and have uncertainties (Kennedy & O'Hagan, 2001). The general practice is to estimate/identify the parameters of a model intended for any purpose. However, we are faced with many different models. These models differ in parametrization, complexity, and performance across various datasets, posing a challenge for model selection. To address this challenge, parameter identification and model selection can be achieved using the Bayesian approach. The Bayesian approach combines our prior beliefs about parameters with data to make conclusions. The Bayesian approach to inference offers a practical advantage by allowing for the easy incorporation of historical or expert knowledge into these models.

1.2 Bayesian inference

The Bayesian paradigm uses Bayes' theorem Eq. (1.1) to update our beliefs based on the new information provided by the data.

$$\underbrace{\pi(\theta|y,M)}_{\text{posterior}} = \frac{\overbrace{f(y|\theta,M)}^{\text{likelihood}}\overbrace{\pi(\theta|M)}^{\text{prior}}}{\underbrace{p(y|M)}_{\text{marginal (averaged) likelihood}}} \tag{1.1}$$

In Bayesian analysis, the prior distribution represents our initial beliefs about the possible values of a parameter before considering the data. The prior, when combined with the likelihood function, which measures the probability of the data given the parameter values, gives the posterior. The posterior distribution is our updated belief of the parameter after incorporating the data.

MacKay (1992) talked of the two levels of inference. Level I involves inferring the parameter values of the model, and level II is model comparison or selection. These levels of inference, illustrated in Fig. 1.1, are sensitive to the choice of priors. Consequently, assessing the impact of priors (Ghaderinezhad et al., 2022) and model selection (Brunetti & Linde, 2018; Brunetti et al., 2017) are active areas of research. *The model selection problem seeks to answer the question: Which model fits the available data well but has few parameters? The prior impact assessment problem seeks to answer the question: How does a prior of interest affect inference?* Prior impact is usually relative to a reference or baseline prior.

There are different types of priors that have been developed and are currently in use. For example, a weakly informative prior provides limited information about the parameter, often allowing the data to play a dominant role in the posterior. In contrast, an uninformative prior provides minimal or no information about the parameter value, allowing the data to shape the posterior distribution completely. Different priors can lead to varying conclusions, highlighting the importance of selecting an appropriate prior based on the context and prior knowledge.

1.3 Model selection

In model selection, we usually seek to find a parsimonious model that provides the same or a similar level of goodness-of-fit as complex models. Model selection techniques balance goodness of fit and complexity (Höge et al., 2018). The goodness of fit can be measured

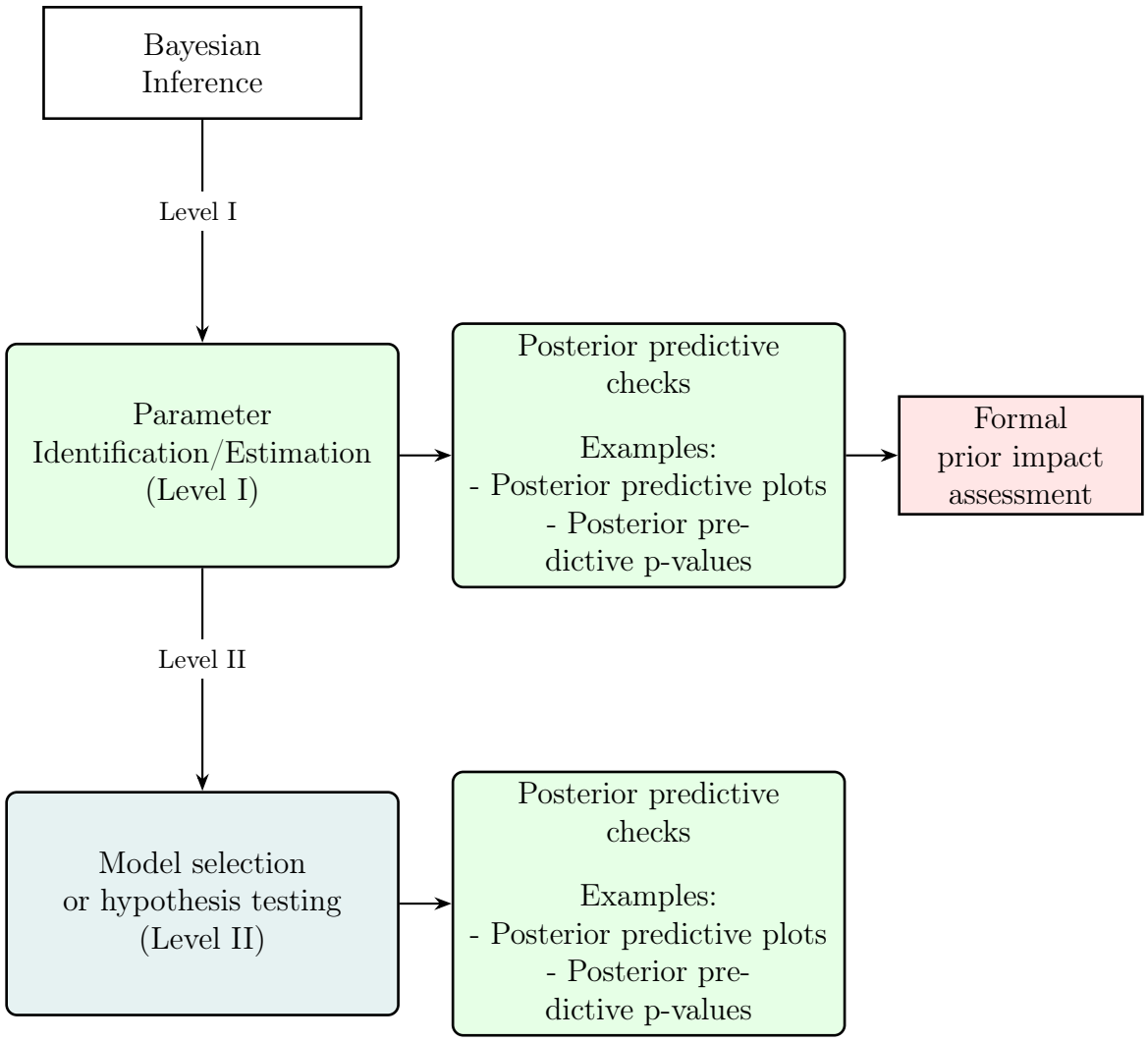


Figure 1.1: The spot of prior impact assessment in Bayesian workflow. This workflow is inspired by that of Gelman et al. (2020). Level I and Level II categories are by MacKay (1992).

by the well-known adjusted R-squared or domain-specific metrics. In this context, the goodness of fit refers to how well the model can reflect the data. Complexity is usually posed in terms of the number of parameters, with the assumption that models vary only in this aspect. This is true for nested models, where a model with more parameters can be reduced into simpler ones by setting one or more parameters to zero. However, complexity is much broader than the number of parameters and includes model structure and computational complexity. For example, hierarchical linear models are more complex than simple linear models due to their multilevel structure. Also, even parametric complexity is not straightforward in the Bayesian context. For models with random parameters like Bayesian and hierarchical models, the effective number of parameters is a more appropriate measure of complexity, as it accounts for the influence of the priors (Spiegelhalter et al., 2002).

By choosing a parsimonious model, we make a trade-off between bias and variance. Complex models tend to have lower bias and higher variance. A biased estimator for model M gives predictions $\hat{y}$ that, on average, are not equal to the observed data y (Geman et al., 1992). The bias of model M is defined as

$$\begin{aligned} \text{Bias}_M &= E(\hat{y}) - y, \\ &= E(\hat{y} - y), \end{aligned}$$

where $E()$ is the expectation. On the other hand, the variance indicates the distance to which individual predictions differ from the mean prediction.

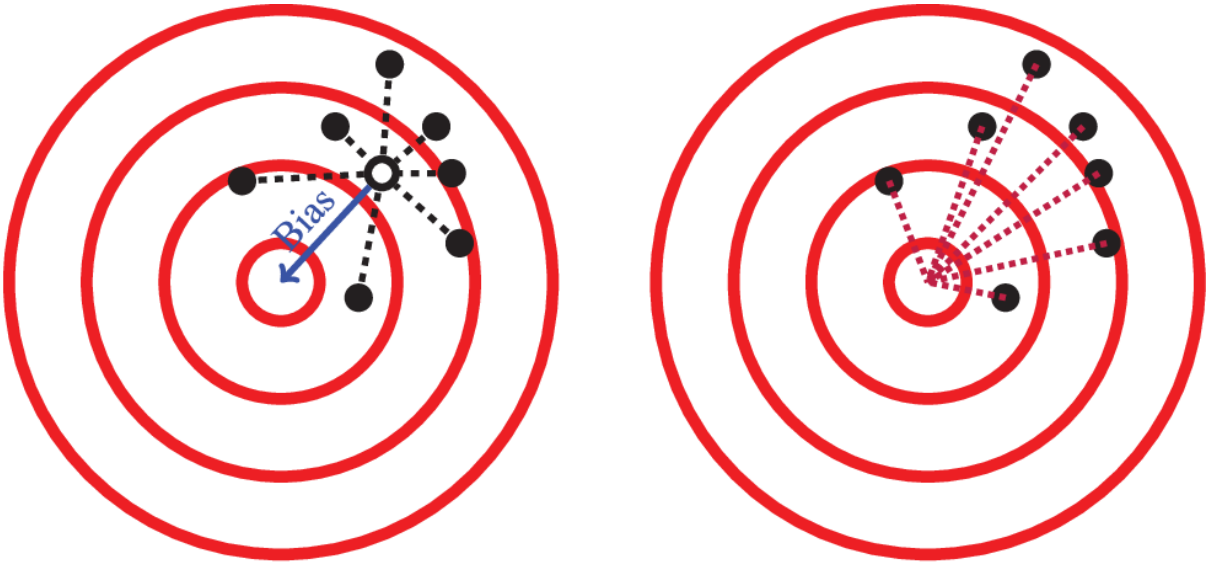
$$\text{Variance} = E[(\hat{y} - E[\hat{y}])^2]$$

The MSE is a function of bias and variance.

$$\begin{aligned} \text{MSE} &= \text{Squared Bias}_M + \text{Variance} \\ &= [E(\hat{y} - y)]^2 + E[(\hat{y} - E[\hat{y}])^2]. \end{aligned}$$

The concept of bias-variance trade-off is illustrated in Fig. 1.2 (Doroudi, 2020). In the game of archery, the centre of the innermost ring gives the highest score, while the outermost ring offers the lowest score. The dots on the rings are markers for the archer's

hits. An archer is said to have a lower bias and higher accuracy if the distance from the centre of the innermost ring to the mean of the black dots (unfilled black circle) is shorter than another archer's. The dashed black lines represent the variance, which are the distances from the centre of the black dots to the individual black dots. An archer is said to have a higher precision if the variance of the dots is lower than another archer's. Bias and variance contribute to the mean squared error, represented by the dashed red lines Fig. 1.2 b. We usually prefer models with low bias and mean squared error (MSE) on the



(a) Bias and variance. (b) Mean-squared error.

Figure 1.2: Depiction of the bias-variance trade-off using an image from Doroudi (2020), licensed under CC BY-NC. The goal is for the points to be as close as possible to the centre of the target (red circle). (a) Bias and variance: The left diagram illustrates the concepts of bias and variance. Bias represents the distance between the average predicted point (marked by the middle dot) and the centre of the target (indicated by the solid blue line). Variance, on the other hand, measures how scattered the individual predictions (black dots) are around the average prediction, as indicated by the dashed black lines. (b) Mean squared error: The right diagram depicts how bias and variance influence the mean squared error (MSE). The MSE combines these two aspects, where the red dashed lines represent the contribution of both bias and variance to the overall error.

training (in-sample) data, which indicates a good fit for the in-sample data. However, models with low bias and high variance that fit in-sample data well may exhibit poor performance on out-of-sample data.The relationship between model complexity, bias, and variance is illustrated in Fig. 1.3. Therefore, the goal is to find an optimal MSE with low bias and variance for in-sample data. For instance, if the in-sample data exhibits a

linear trend, a linear model will achieve low bias and low variance, fitting the data well. Conversely, while potentially fitting the in-sample data with similarly low bias, a quintic model may introduce higher variance due to its increased flexibility. This higher variance means the quintic model is more sensitive to fluctuations in the in-sample data and may not perform as well with new, unseen data. Thus, the challenge is selecting a model that fits the in-sample and out-of-sample data well.

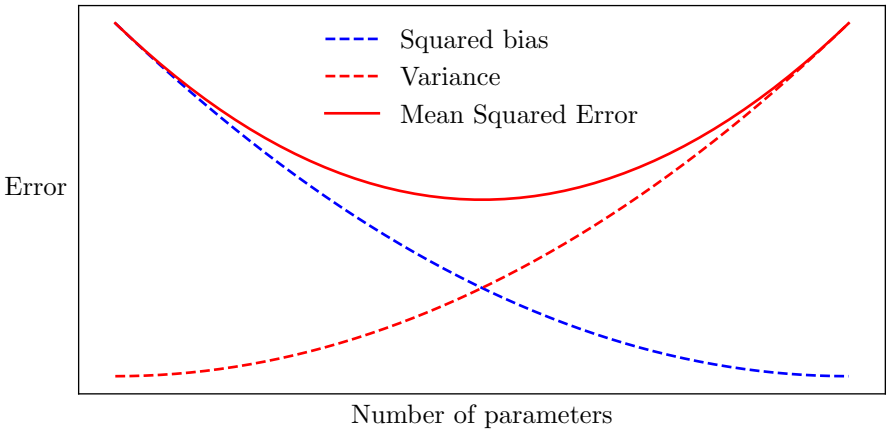


Figure 1.3: Relationship between model complexity measured by the number of parameters, bias and variance. It is assumed that the models differ only in the number of parameters. The optimal mean squared error is where the bias and the variance intersect. Variance is the error introduced by a model being too sensitive to small fluctuations in the in-sample data. When the number of model parameters increases, the model can fit the in-sample data very closely, but this can lead to poor predictions on new data. This results in higher variance (more error). For examples with different types of models see Geman et al. (1992).

Indeed, model selection methods penalise complexity to reduce model variability while considering bias or accuracy. To highlight the importance of model selection, we will demonstrate it with polynomial models. We generate data from a 3-degree polynomial as follows.

$$
\begin{aligned}
&y = 5x - x^2 + 0.8x^3 + \epsilon, \\
&x = (x_1, \ldots, x_n), \quad n = 100, \\
&\epsilon \sim \mathcal{N}(0, 3.5).
\end{aligned}
\tag{1.2}
$$

First, we split our data (y) into two parts: the first 75%, called in-sample data, is used to calibrate the model using the ordinary least squares method. Once calibrated, we use the model to predict the remaining 25% of the data, known as out-of-sample data. The results plotted in Fig. 1.4 are as follows: the linear model does not fit the in-sample data (blue) and the out-of-sample data (red) well Fig. 1.4. On the other hand, quadratic, quintic, and 30-degree polynomial models fit the in-sample data well but perform poorly on out-of-sample data, which suggests overfitting. The cubic model, our data-generating model, fits both the in-sample and out-sample data well, even though the latter is not used for model calibration. The 30-degree polynomial model, due to its higher complexity, exhibits a greater range of predictions, indicating increased variance. The goodness of fit is measured by the adjusted R^2, where 1.00 is the best possible value and negative values indicate an inability to predict out-of-sample data. Model selection would be challenging based on the adjusted R^2 for in-sample data since the higher-degree polynomial models all share close adjusted R^2. Thus, we need a model selection technique. An example is provided in Section 2.3.2 where all models fit the data equally well.

When we use the entire dataset to calibrate the model and then predict the last 25% of observations (double use of data), we observe similar accuracy in the predictions measured by the R^2 across different polynomial models (3, 5, 30) as depicted in Fig. 1.5. However, in practical situations, future data is not available. This underscores the importance of selecting models carefully as we rely on them to forecast future outcomes.

1.3.1 Implicit and explicit penalty in model selection

Model selection techniques attempt to balance bias and variance. Models with many parameters usually fit current or in-sample data well but have a higher error on future data, or in machine learning terms, they fit training data well but have a lower fit on test and validation data. Model selection techniques penalise the number of parameters, as models with many parameters have lower bias and hence are a good fit on training data but usually have lower fit on test or validation data. The aim is to choose models with fewer parameters but equally good fit for training data as models with many parameters. Also, simpler models require fewer computing resources. The search for simpler models is

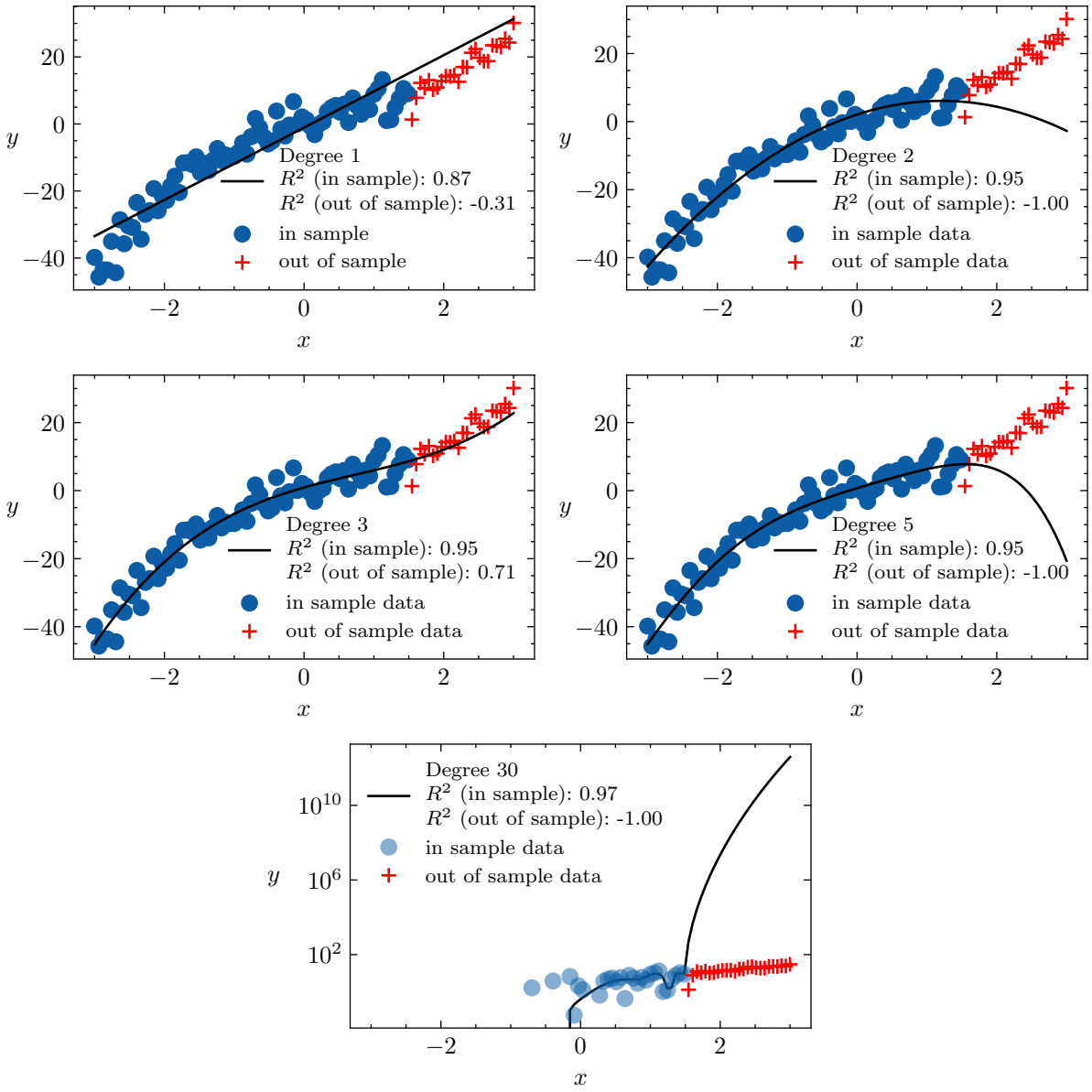


Figure 1.4: Illustration of model complexity on goodness of fit, measured by the adjusted R^2, for in-sample and out-of-sample data. As the degree of the polynomial increases, the model becomes more complex and fits the in-sample data better. However, beyond a certain point (degree 5 in this case), increased complexity leads to overfitting and poor performance on out-of-sample data. Higher degree polynomial models (2, 5, and 30) with more parameters have higher goodness of fit for in-sample data but poor goodness of fit for out-of-sample data. The cubic polynomial model is the best as it balances the bias and variance, leading to a high goodness of fit for in-sample and out-of-sample data. The 30-degree polynomial model is too wiggly, meaning it has captured the noise. Hence, it also fails to have a high goodness of fit for the out-of-sample data.

not limited to dynamic modelling. For example, all the large language model developers are now looking for ways to get the same model accuracy with as few parameters as pos-

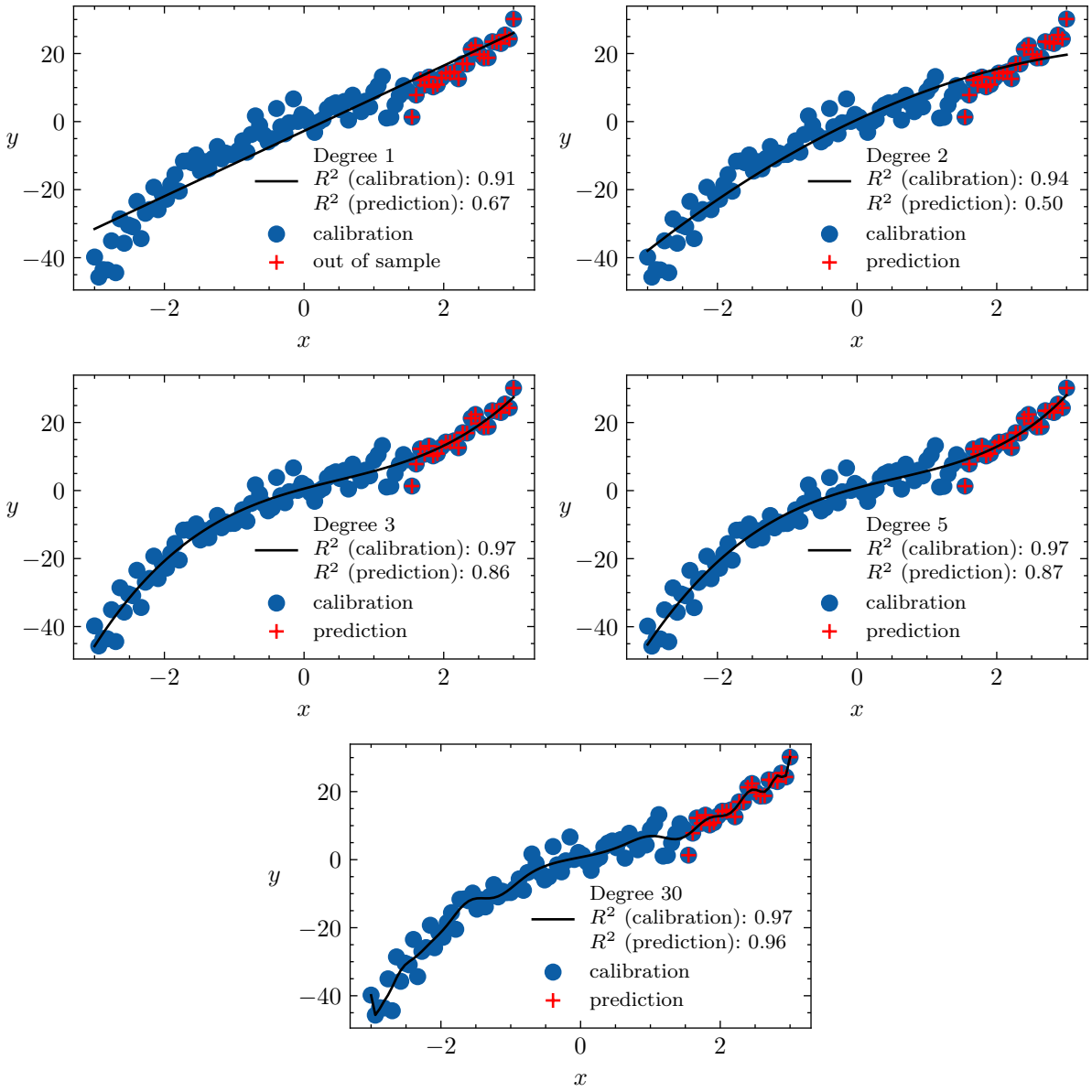


Figure 1.5: Illustration of model complexity on goodness of fit, measured by the adjusted R^2. Here, the entire dataset is used to calibrate the models. Subsequently, the calibrated models are used to predict 25% of the data, which was also used in the calibration (double use of data). We see that all the models have done better in predicting the last 25% of the data than in Fig. 1.4. Also, the cubic, quintic and 30-degree polynomial models all have comparable R^2.

sible. Model selection techniques introduce a penalty term for the number of parameters such that a complex model gets a higher penalty. This penalty can either be explicit, as in the case of information-theoretic approaches like information criteria, or implicit, as in the case of the Bayes factor. In the latter, there is implicit penalisation through the model evidence. The model evidence, denoted as z, can be defined as follows (Llorente

et al., 2023)

$$Z = \ell_{\max} W, \quad W \in [0, 1], \tag{1.3}$$

where

$$W = \frac{1}{\ell_{\max}} \int p(\theta) p(y|\theta) d\theta,$$

$\ell_{\max}$ is the maximum value of the likelihood function, $p(y|\theta)$ is the likelihood function and $p(\theta)$ is the prior. The prior and likelihood function are discussed in detail in other chapters.

The term W is an implicit penalty that depends on the prior and the sample size. The idea of implicit penalisation can be traced back to MacKay (2003). The log evidence is

$$\log(Z) = \log \ell_{max} + \log W, \tag{1.4}$$

where $\log \ell_{\max}$ is a goodness of fit measure and $\log W$ is the penalty term. This is similar to information-theoretic-based approaches, where goodness of fit is penalised based on the number of model parameters. The evidence is also known as the marginal likelihood. The marginal likelihood is used in Bayesian model comparison and can also be used to weight various models in Bayesian model averaging.

One approach for choosing between models from a predetermined set is the BF (Berger & Pericchi, 1996). The strength of the BF lies in its ability to balance model complexity and goodness-of-fit with minimal assumptions. However, calculation of the marginal likelihood (the denominator in Bayes' theorem), which is necessary for the BF, is challenging and costly due to the need for multiple model runs, the potential for multimodal and highly correlated posteriors, and the inherent complexity of the integration problem. Due to these challenges, model uncertainty is often evaluated using information-theoretic criteria (Birgé & Massart, 2007; Bai et al., 1999) that explicitly penalise model complexity based on the number of parameters and under limiting assumptions (see e.g., Berger et al., 2001).

There are various methods for numerically estimating the BF. Llorente et al. (2023) provided a comprehensive review of commonly used techniques for computing the marginal likelihood and the BF. These methods include naive Monte Carlo, the harmonic mean

estimator (Newton & Raftery, 1994), the generalized harmonic mean estimator (Gelfand & Dey, 1994), importance sampling and Chib's method (Chib & Jeliazkov, 2001; Chib, 1995), bridge sampling (Meng & Wong, 1996; Gelman & Meng, 1998), nested sampling (Skilling, 2004, 2006), and thermodynamic integration (Calderhead & Girolami, 2009; Lartillot & Philippe, 2006; Ogata, 1989), which is the method used in our study. Thermodynamic integration is well-suited for high-dimensional integrals (Ogata, 1989, 1990) involving physics-based models, such as ordinary differential equation (ODE) systems. Naive Monte Carlo methods are generally unstable and inefficient, requiring many samples for convergence. Choosing suitable distributions is crucial for importance sampling and harmonic mean estimators. The effectiveness of bridge sampling also relies on selecting a good proposal distribution, which can be challenging to determine in advance as it has to overlap with the posterior distribution. One of the main difficulties with nested sampling is generating samples from a truncated prior as the threshold increases. Chib's method's efficiency is influenced by how close an arbitrary value is to the posterior mode.

Turning to hydrology, multiple studies (e.g. J. Zhang et al., 2020; Brunetti et al., 2017; Zheng & Han, 2016; Shafii et al., 2014; Laloy & Vrugt, 2012) use a gradient-free algorithm, differential evolution adaptive Metropolis (DREAM) (Vrugt, 2016) for posterior parameter inference and to compute the marginal likelihood.

The DREAM algorithm has been designed with an acceptance rate comparable to that of the random walk Metropolis (RWM) algorithm, which has an optimal acceptance rate of 0.234 (Vrugt et al., 2008; Gelman, Roberts, & Gilks, 1996; Roberts & Rosenthal, 2009). The acceptance rate or probability refers to the fraction of proposed samples that are accepted. In contrast, a gradient-based sampler like HMC generally has a much higher optimal acceptance rate, about 0.65 (Radford M. Neal, 2011; Beskos et al., 2013).

The model selection part of this thesis is closely related to those of Brunetti et al. (2019, 2017); Brunetti & Linde (2018) who computed Bayes factors for conceptual hydrogeological models.

1.4 Ordinary differential equations

Continuous dynamical systems can be described by ordinary differential equations. A general ODE can be expressed as

$$\frac{\mathrm{d}z}{\mathrm{d}t} = F(t, z, \theta), \tag{1.5a}$$

$$\hat{z} = z(t = 0), \tag{1.5b}$$

where the state $z(t)$ evolves with time $t \in (0, T]$. The initial condition is $\hat{z} \in \mathbb{R}$, and $F : (0, T] \times \mathbb{R} \times \mathbb{R}^p \to \mathbb{R}$ is a known function specific to the dynamic system. The ODE in Eqs. (1.5a) to (1.5b) usually has no closed form, so numerical methods are used to estimate the parameters. The Bayesian approach is used to estimate the parameters and obtain uncertainty bands. Ordinary differential equations (ODEs) are used in various fields, such as hydrology, epidemiology, ecology, and engineering.

1.4.1 Hydrological models

Hydrology is the study of the interaction between water and the environment. It plays a major role in sectors such as hydropower. For example, the Hydrologiska Byråns Vattenbalansavdelning (HBV) model (Seibert & Bergström, 2022) was developed to forecast reservoir inflow for the hydropower sector and give flood warnings (Bergström & Lindström, 2015). Early warnings for weather events such as flash floods can mitigate the effects. With the increasing demand for water due to population growth, agriculture, and manufacturing industries, hydrological models are essential for sustainable water resource management (Devia et al., 2015; Zalewski, 2002). Hydrology has interesting research questions and observed real-world data to which our methods can be applied. Some of these questions include: *How do vegetation and climate affect discharge? How extreme can the discharge get? How much water is currently available and will be available in the future? What is the average discharge at the minimum and maximum flow?* (Seibert & Bergström, 2022). Various types of models have been developed to answer these questions. These models vary in structure, parametrization, and performance across different catchments. Therefore, hydrologists are faced with the task of model selection. The aim

is to select a hydrological model that best describes current data with fewer parameters.

A commonly used type of model to represent hydrological processes is the conceptual model. A conceptual rainfall-runoff model uses fluxes to interconnect various soil layers, thereby describing the movement of water between these layers (Duan et al., 1992). The main components of most conceptual models are a reservoir for storage and release of water, a lag function for transmission and delay of fluxes and an element for combining, splitting and or scaling of fluxes (Fenicia et al., 2011). Some examples include the topography-based hydrological model (TOPMODEL) (BEVEN, 1997), variable infiltration capacity (VIC) (Wood et al., 1992) and the HBV. The movement of water between layers can be represented by ODEs. For instance, Fenicia et al. (2011) represent the flow of water between storage compartments using ODEs. The performance of these conceptual models is dependent on the physical characteristics (Kavetski & Fenicia, 2011) and scale (Merz et al., 2009) of various catchments. We focus on the HBV model which describes the movement of water from the atmosphere in different forms to the soil, rivers, and lakes. Some of the water is taken up by plants and lost through evapotranspiration. We exemplify our approach to model selection in hydrology. There are several variants of the HBV model (Jansen et al., 2021). We decided to mimic the HBV model to illustrate our approach to model selection. We developed several ODEs of varying complexities. The ODE models maintain the main ideas in the HBV but do not have a snow routine.

1.4.2 Susceptible Exposed Infectious Recovered

One commonly used model to study the dynamics of a disease in a population is the SEIR. The SEIR model is a type of compartmental model that divides the population into different compartments. The S compartment includes individuals who are susceptible to the disease, the E compartment is for individuals who are exposed to the disease but cannot yet infect others, the I compartment represents individuals currently infected and able to infect others, while the R compartment is for individuals who have recovered from the disease. The SEIR model is a system of ordinary differential equations that describe how a disease spreads in a population. In some cases, the total population changes over time (Li et al., 1999). In other instances, the total population is assumed to be constant

(Li & Muldowney, 1995).

The SEIR model has been used to model contagious diseases such as the Ebola outbreak (Frasso & Lambert, 2016) and vector-borne diseases such as malaria (Shah & Gupta, 2013). The model has been widely used to study the spread of COVID-19 (He et al., 2020; Kemp et al., 2021; Kamrujjaman et al., 2022). This SEIR model can been extended to include other compartments. For instance, the study by Mwalili et al. (2020) examines the spread of COVID-19 while accounting for social distancing by introducing an additional compartment called pathogens (P), resulting in the SEIR-P model. A deceased (D) compartment is sometimes added (Davarci et al., 2024; Quintero et al., 2021) leading to the SEIRD model. Other compartments, such as intensive care unit (ICU) and hospital (H), have been included in several studies on COVID-19 (Shah & Gupta, 2013).

The Bayesian approach is widely employed to estimate the parameters of the SEIR model and make predictions. For example, Lai et al. (2021) employed the Bayesian approach to estimate the parameters of the SEIR model for COVID-19 on the Diamond Princess Cruise Ship. Also, using the Bayesian approach, Yin et al. (2022) estimated the parameters of an extended SEIR model for the first wave of COVID-19 in India. However, different priors are used for the models without any formal prior impact assessment.

1.5 Prior impact assessment

Bayesian methods are commonly employed to solve parameter identification problems of continuous-time dynamical systems described by ODEs (Girolami, 2008). These models usually involve unknown parameters that significantly influence the system's behaviour. Therefore, it is crucial to identify these parameters using observed data before utilising the model.

Parameter identification problems for ODEs usually involve sparse data since observations are available only at specific time points. Also, there are usually only observations on some states. Due to the sparse nature of the data, it is often difficult to fully constrain the parameters, which can lead to inherently challenging inference problems. In addition to the data sparsity issue, the parameters are typically constrained by physical considerations, such as positivity or expert knowledge.

As a result of these issues, Bayesian methods are well suited for parameter identification problems involving ODEs. Using a prior can transform an ill-conditioned problem caused by limited data into an identifiable problem. Non-informative priors are rarely used for ODEs. Therefore, there is a need to assess how the choice of prior impacts posterior inference. For instance, do the priors have a similar impact, or does one have a lower impact on the posterior summary statistics?

A common practice is to refit the model with slightly different priors and check if the posterior summary statistics change (Stefan et al., 2022; Pedroza et al., 2018; Schmidli et al., 2014; Nur et al., 2009). This approach provides a qualitative understanding of how different priors affect posterior statistics but usually does not offer a quantitative measure of the differences between the resulting posteriors. Some studies have employed divergences to do a formal prior impact assessment. Tang et al. (2016) have used the Kullback–Leibler (KL) and Weiss (1996) the chi-squared divergence. Others have used proper metrics for prior impact assessment. Roos et al. (2015) used the Helinger distance. In addition, the Wasserstein distance has been used (Jones et al., 2022; Ghaderinezhad et al., 2022). All these quantitative measures have been employed in the context of Generalized linear models (GLMs) or simple distributions. GLMs are models where the response variable can be expressed as coming from an exponential family of distributions. Examples of GLMs are simple linear, logistic, and poison regression. These measures have an interpretation problem, making the area of prior impact assessment an active research. For instance, a value of 0.5 has no intuitive interpretation, such as a higher or lower impact.

Formal prior impact assessment can be performed after level I Bayesian analysis as shown in Fig. 1.1. Prior impact assessment is absent for systems modelled by ODEs. We build upon the WIM introduced by Ghaderinezhad et al. (2022) to perform prior impact assessment for ODEs.

1.6 Contributions

This thesis introduces new methodologies for Bayesian model selection and prior impact assessment, with an emphasis on dynamical systems. It also offers insights into

the prior-to-posterior transition in Bayesian analysis. Several examples are provided to demonstrate the application of these methods. The main contributions are summarised below.

1. Several methods exist to compute the marginal likelihood, hence the Bayes factor (Llorente et al., 2023), for the full Bayesian model selection. These methods have some disadvantages. For instance, for methods that require a proposal distribution, the accuracy depends on how well the proposal distribution overlaps with the posterior distribution. These methods include bridge sampling (Meng & Wong, 1996; Gelman & Meng, 1998), importance sampling, and harmonic mean estimators (Gelfand & Dey, 1994). Thus, the accuracy of these methods depends on the choice of a proposal distribution. The most used method is TI (Calderhead & Girolami, 2009; Lartillot & Philippe, 2006; Ogata, 1989, 1990). However, sampling-based methods are used to get posterior samples used for TI. These sampling-based methods sometimes do not scale well and require more time to converge compared to gradient-based methods (Radford M. Neal, 2011). Also, in the case of multimodal posteriors, the methods used might not explore the entire posterior. Hence, by taking advantage of differentiable software, we introduce a new gradient-based algorithm, Replica exchange Hamiltonian Monte Carlo (REHMC), to sample across multimodal posterior and scales well. REHMC + TI is used to estimate the marginal likelihood, and we demonstrate its effectiveness with real and synthetic rainfall/runoff data. This thesis advocates using differentiable programming languages to implement hydrological models to improve computational efficiency and motivate hydrologists to adopt these tools for conceptual model development.

2. The prior is a critical component of any Bayesian analysis and impacts inference. Hence, Bayesian analysis usually involves some assessment of the prior. The assessment can be qualitative, which consists of trying different priors to see if the results change based on the prior used (Stefan et al., 2022; Pedroza et al., 2018; Schmidli et al., 2014). Also, there are several approaches to formally assessing the impacts of priors (Morita et al., 2008; Wiesenfarth & Calderazzo, 2020; Jones et al., 2022). However, these approaches do not have a clear interpretation in the sense

that it is difficult to assess the magnitude of the impact. Furthermore, prior impact assessment has not been applied to ODEs even though the Bayesian approach is widely used for ODEs (Girolami, 2008). In this thesis, the WIM proposed by Ghaderinezhad et al. (2022) is extended to Bayesian models involving ODEs by using advances in computational optimal transport (Cuturi, 2013; Cuturi et al., 2022; Cuturi & Doucet, 2014). A new prior impact measure, known as the prior scaled Wasserstein Impact Measure (sWIM) that improves on the WIM by endowing it with a relative sense of scale is introduced making it easier to interpret the impact of the prior on the posterior inference.

3. There are studies on how priors influence posteriors (Roos & Held, 2011; Roos et al., 2015; Morita et al., 2008; Wiesenfarth & Calderazzo, 2020; Jones et al., 2022). However, there is limited understanding of how priors evolve to the posteriors. To investigate how priors transition to posteriors, we employ power posteriors (Friel & Pettitt, 2008) introduced in Chapter 2 in the context of model selection and Wasserstein distances introduced in Chapter 3. Power posteriors construct a continuous path from the prior to the standard posterior (Gelman & Meng, 1998). We use power posteriors and Wasserstein distances to gain insights into prior to posterior transitions. Our findings reveal that the distance between power posteriors and priors stabilises at a power value lower than 1.0. We have also demonstrated that subsampling is equivalent to using power posteriors and derived new analytic expressions for power posteriors of conjugate cases. This indicates there is a subsample size that contains the same information as the full sample size. In some studies, it might be impractical to have large sample sizes and smaller sample sizes are needed. An example of such studies is clinical trials for rare diseases. It is standard practice to do interim analysis at predetermined times and stop if the sample size has enough information (Berry, 2006; Ryan et al., 2022). Consequently, we introduce the concept of saturated sample size to indicate when a small sample size can give the same information as a large sample size. The saturated sample size can be calculated as new data is collected.

1.7 Thesis outline

This chapter introduces Bayesian inference, model selection, and prior impact impact assessment. It emphasises the importance of model selection through illustrative synthetic examples. There is also an introduction to modelling continuous dynamical systems.

In Chapter 2, we use a gradient-based sampler to compute the model evidence necessary for full Bayesian model selection. We demonstrate that this approach accurately calculates the model evidence through examples with known evidence. We then show the BF can decisively favour a data-generating model when information-theoretic-based approaches like DIC and WAIC fail to do so. Furthermore, we illustrate our method using real-world hydrology data for ODE based models. All this is made possible by differentiable programming that utilises automatic differentiation. Differentiable software uses automatic differentiation to obtain the derivative of any function that would otherwise be difficult without an analytic solution. As part of open science, our codes are available in public repositories.

Chapter 3 extends Bayesian prior impact assessment to ODEs. We build on the literature of Wasserstein distances from statistics and optimal transport to introduce a new interpretable prior impact assessment measure known as sWIM. Our method is exemplified using real-world predator-prey data and COVID-19 first-wave data for Luxembourg. Again, our codes are available in public repositories.

Chapter 4 investigates how priors evolve into posteriors by employing power posteriors and Wasserstein distances. Power posteriors make a continuous path from the prior to the standard posterior. By building on power posteriors, Chapter 4 introduces the concept of saturation sample size, a smaller sample size that provides information similar to a larger sample size. Also, it shows that subsampling is equivalent to using power posteriors and provides analytic expressions for power posteriors of conjugate cases.

Finally, Chapter 5 presents the conclusions and future work, summarising the key findings and suggesting further research directions.

1.8 Research dissemination

Article

- **Mingo, D. N.**, Nijzink, R., Ley, C., & Hale, J. S. (2024). Selecting a conceptual hydrological model using Bayes' factors computed with Replica Exchange Hamiltonian Monte Carlo and Thermodynamic Integration. Geoscientific Model Development., 18, 1709–1736, https://doi.org/10.5194/gmd-18-1709-2025, 2025

Manuscript under review

- **Mingo, D. N.**, Hale, J. S., & Ley, C. (2024). Bayesian prior impact assessment for dynamical systems described by ordinary differential equations. University of Luxembourg Open Repository and Bibliography [Submitted preprint]. `https://hdl.handle.net/10993/61471`

Manuscript in preparation

- **Mingo, D. N.**, & Hale, J. S. (2024). Insights into the prior to posterior transition through Wasserstein distances and the power posterior [Manuscript in preparation].

Conferences

- **Mingo, D. N.**, Ley, C., & Hale, J. S. (2023, December 16). Using optimal transport to assess the impact of prior choice on Bayesian parameter inference in dynamical systems. *16th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics)*, Berlin, Germany. `https://www.cmstatistics.org/RegistrationsV2/CMStatistics2023/viewSubmission.php?in=1551&token=32739r7102991s97nr1p3oqn479r71q1`

- **Mingo, D. N.**, Nijzink, R., Ley, C., Schymanski, S., & Hale, J. S. (2023, April 24). Thermodynamic integration via replica exchange Hamiltonian Monte Carlo for

faster sampling and model comparison. *EGU General Assembly*, Vienna, Austria. https://doi.org/10.5194/egusphere-egu23-2910

- **Mingo, D. N.**, Nijzink, R., Ley, C., Schymanski, S., & Hale, J. S. (2023, January 25). Using replica exchange Hamiltonian Monte Carlo and thermodynamic integration for comparison of dynamic rainfall-runoff models. *Luxembourg-Waseda Conference on Modelling and Inference for Complex Data*, Belval, Luxembourg. https://hdl.handle.net/10993/54843

- **Mingo, D. N.**, Nijzink, R., Ley, C., Schymanski, S., & Hale, J. S. (2022, June 5). Using Bayes factors to compare dynamical models of hydrological systems. *5th International Conference on Econometrics and Statistics (EcoSta 2022)*, Online and Kyoto, Japan. https://hdl.handle.net/10993/51573

Chapter 2

Model selection using Bayes' factors computed with Replica Exchange Hamiltonian Monte Carlo

The content of this chapter is based on the article.

Mingo, D. N., Nijzink, R., Ley, C., and Hale, J. S. (2024). Selecting a conceptual hydrological model using Bayes' factors computed with Replica Exchange Hamiltonian Monte Carlo and Thermodynamic Integration. Geoscientific Model Development., 18, 1709–1736, https://doi.org/10.5194/gmd-18-1709-2025, 2025

Author contributions.

DNM: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualisation, Writing - original draft, Writing - review & editing. RN: Conceptualization, Formal analysis, Methodology, Writing - review & editing. CL: Conceptualization, Funding acquisition, Formal analysis, Methodology, Supervision, Writing - review & editing. JSH: Conceptualization, Formal analysis, Funding acquisition, Methodology, Project administration, Software, Supervision, Validation, Writing - original draft, Writing - review & editing.

Abstract

We develop a method for computing Bayes' factors of conceptual rainfall-runoff models based on thermodynamic integration, gradient-based replica-exchange Markov Chain Monte Carlo algorithms and modern differentiable programming languages. We apply our approach to the problem of choosing from a set of conceptual bucket-type models with increasing dynamical complexity calibrated against both synthetically generated and real runoff data from Magela Creek, Australia. We show that using the proposed methodology the Bayes factor can be used to select a parsimonious model and can be computed robustly in a few hours on modern computing hardware.

2.1 Introduction

Hydrologists are often faced with assessing the performance of models that differ in their complexity and ability to reproduce observed data. The Bayes factor (BF) is one method for selecting between models from an *a priori* chosen set (Berger & Pericchi, 1996). The appeal of the BF lies in its ability to implicitly and automatically balance model complexity and goodness-of-fit under few simplifying assumptions. The BF is also invariant to data and parameter transformations unlike information theory-based criteria such as Akaike information criteria (AIC) and Bayesian information criterion (BIC) (O'Hagan, 1997). For example, a logarithmic transformation of the discharge or the square root of a parameter such as the flow rate can accelerate the convergence of the model, but it will not affect the computed BF.

However, the BF requires the computation of the marginal likelihood (the denominator in Bayes theorem) for each model, which is a difficult and expensive integration problem. This expense and difficulty can be attributed to three main factors; the necessity of many model runs at different points in the parametric space; the possibly multi-modal and highly correlated nature of the posterior that can lead to isolated and/or slowly mixing chains; and finally the inherent difficulty of the marginal likelihood integration problem.

Because of these difficulties, it is the case today that the BF is not widely used by practitioners, despite it being a crucial component in Bayesian model comparison, selection and averaging (Höge et al., 2019). This stands in contrast with the widely studied and used Bayesian parameter estimation procedure (Gelman et al., 2020). Consequently, model uncertainty is often ignored, or assessed by either *ad hoc* techniques or information theoretic criteria (Birgé & Mas-

sart, 2007; Bai et al., 1999) that explicitly (rather than implicitly) penalise model complexity based on some measure of the number of parameters and under limiting assumptions, see e.g. (Berger et al., 2001) for a full discussion.

2.1.1 Background

Looking outside of hydrology, there are a number of notable works that have developed techniques for numerically estimating the BF. A recent comprehensive review by Llorente et al. (2023) discusses the relative advantages of commonly used methods for computing the marginal likelihood, and consequently, the BF, such as naive Monte Carlo methods, harmonic mean estimator (Newton & Raftery, 1994), generalised harmonic mean estimator (Gelfand & Dey, 1994), importance sampling and Chib's method (Chib & Jeliazkov, 2001; Chib, 1995), bridge sampling (Meng & Wong, 1996; Gelman & Meng, 1998), nested sampling (Skilling, 2004, 2006) and finally thermodynamic integration (Calderhead & Girolami, 2009; Lartillot & Philippe, 2006; Ogata, 1989), the technique that we choose to use in this study. Thermodynamic integration is well suited for high dimensional integrals (Ogata, 1989, 1990) involving physics-based models such as Ordinary differential equation (ODE) systems. The naive Monte Carlo is unstable and usually not efficient, requiring a huge number of samples for convergence. The importance sampling and harmonic estimators require a suitable choice of the importance and proposal distributions, respectively. The performance of bridge sampling also depends on a good choice of proposal distribution, which in practice is not straightforward to determine *a priori*. The main difficulty with nested sampling is generating samples from a truncated prior as the threshold increases (Chopin & Robert, 2010; Henderson & Goggans, 2019). However, the efficiency of Chib's method depends on how close an arbitrary value is to the posterior mode (Dai & Liu, 2022). Hug et al. (2016) illustrated that Chib's method significantly underestimates the marginal likelihood of a bimodal Gaussian mixture model.

Turning our attention to works within hydrology that develop methods for computing Bayes factors, to the best of our knowledge, the seminal work by Marshall et al. (2005) was the first to propose computing Bayes factors for hydrological model selection. Marshall et al. (2005) used Chib's method to estimate the marginal likelihood of conceptual models. More recently various other authors (P. Liu et al., 2016; Brunetti et al., 2019, 2017; Volpi et al., 2017; Cao et al., 2019; Brunetti & Linde, 2018; Marshall et al., 2005) have considered the computation of Bayes factors in a hydrological or hydrogeological context.

Perhaps most closely related to our study are the recent works of Brunetti et al. (2019, 2017); Brunetti & Linde (2018) who computed Bayes factors for conceptual hydrogeological models with thermodynamic integration techniques. Brunetti et al. (2017) compared naive Monte Carlo, bridge sampling based on the proposal distribution developed by Volpi et al. (2017), and the Laplace Metropolis method in terms of calculating the marginal likelihood of conceptual models. Like most studies, the naive Monte Carlo approach performed poorly. Also, Brunetti & Linde (2018) computed the marginal likelihood using methods based on a proposal distribution, notably bridge sampling. Several marginal likelihood estimation methods have been compared within hydrological studies. For example, P. Liu et al. (2016) found that thermodynamic integration gives consistent results compared to nested sampling and is less biased.

Many studies in hydrology, e.g. J. Zhang et al. (2020); Brunetti et al. (2017); Zheng & Han (2016); Shafii et al. (2014); Laloy & Vrugt (2012) and Kavetski & Clark (2011) have used the differential evolution adaptive Metropolis (DREAM) algorithm (Vrugt, 2016) for posterior parameter inference. Volpi et al. (2017) introduced a method to construct the proposal distribution for bridge sampling and integrated it with the DREAM algorithm. However, it still requires the user to choose the number of Gaussian distributions for the Gaussian mixture proposal distribution. The DREAM algorithm has been developed with an acceptance rate similar to the RWM algorithm, which has an optimal acceptance rate of 0.234 (Vrugt et al., 2008; Gelman, Roberts, & Gilks, 1996; Roberts & Rosenthal, 2009). The acceptance rate or probability is the proportion of the proposed samples accepted in the Metropolis-Hastings algorithm. In contrast, a gradient-based sampler such as HMC, which we use in this work, typically has a far higher optimal acceptance rate of around 0.65 (Radford M. Neal, 2011; Beskos et al., 2013). In addition, gradient-based samplers show improved chain mixing properties in high dimensions and on posteriors with strongly correlated parameters (Radford M. Neal, 2011). Gradient-based algorithms have been used in hydrology for parameter estimation, but not model selection. For instance, Hanbing Xu & Guo (2023) found that NUTS sampler (Hoffman & Gelman, 2014) performed well for calibrating a model of daily runoff predictions of the Yellow River basin in China. Krapu & Borsuk (2022) employed HMC to calibrate the parameters of rainfall-runoff models. The model selection studies by P. Liu et al. (2016) and Brunetti et al. (2017, 2019) that use the BF use posterior samples from the DREAM algorithm, and consequently a lower acceptance rate than gradient-based samples e.g. HMC. In addition, because gradient-based samplers incorporate information about the local geometry of the posterior, they are usually easier to tune to achieve the optimal acceptance rate, particularly in the moderate or high-dimensional parameter setting

(num. parameters > 5).

2.1.2 Contributions

The overall contribution of this paper is to describe the development of a method, REpHMC, which, when used in conjunction with thermodynamic integration (TI), can be used to estimate the BF of competing conceptual rainfall-runoff hydrological models. Our approach for estimating the marginal likelihood combines TI for marginal likelihood estimation, REMC for power posterior ensemble simulation and pHMC for high-efficient gradient-based sampling which in sum we call the REpHMC + TI estimator. We demonstrate that REpHMC can sample from moderate-dimensional, strongly correlated and/or multimodal distributions that frequently arise from hydrological models. In addition, REpHMC + TI can obtain posterior parameter estimates and the marginal likelihood simultaneously. We remark that Brunetti et al. (2019) also suggested, but did not explore, the idea of using REMC (therein called parallel tempering Monte Carlo) to improve chain mixing in hydrological models. Two other gradient-based samplers, MALA (Xifara et al., 2014) and NUTS (Hoffman & Gelman, 2014) are used briefly in this paper as a point of comparison, but we do not include their detailed derivation.

Another key contribution of our work compared with e.g. Brunetti et al. (2017, 2019) is the incorporation of recent ideas from probabilistic programming for the automatic specification of the Bayesian inference problems (parameter and BF estimation). Utilising recent techniques from the literature on neural ordinary differential equations (ODEs) (R. T. Chen et al., 2018; Rackauckas et al., 2020; Kelly et al., 2020), we formulate a set of Hydrologiska Byråns Vattenbalansavdelning (HBV)-like models with extensible model complexity as a system of Ordinary differential equations (ODEs). Working in this framework allows us to use efficient high-order timestepping schemes for the numerical solution of the ODE system and to automatically derive the associated continuous adjoint ODE system. With this adjoint system we can efficiently calculate the derivative of the posterior functional with respect to the model parameters, a necessary step for working with gradient-based samplers such as HMC. We emphasise at this point that our approach is largely free of manual tuning parameters and straightforward to implement in a differentiable programming framework (we use TensorFlow probability (TFP) with the JAX backend, but the ideas are applicable in similar frameworks such as Stan or PyMC3). We remark that a recent more theory-focused paper (Henderson & Goggans, 2019) also proposed using TI with HMC via the Stan probabilistic programming language, but with results for non-time series

models and without using REMC, which is an important aspect of our approach.

After model selection via the BF, it is essential to check if the chosen model can generate the observed data. Hydrographs show the time series of stream flow. However, formal goodness-of-fit testing is necessary since it is challenging to see a mismatch in hydrographs for dense data. We therefore use the prior calibrated posterior predictive p-value (PCPPP), which simultaneously tests for prior data conflict and discrepancies in the model for further improvements.

In summary, this paper is the first to propose the REpHMC + TI method in a probabilistic programming framework for the estimation of marginal likelihoods related to hydrological systems in view of model selection. We demonstrate the performance of our method by showing a) a validation of the methodology using an analytically tractable model, b) its improved efficiency with respect to classical methods using artificially generated data, and c) an application of a Bayes factor based model selection on real rainfall/runoff data collected from the Magela Creek catchment in Australia.

Our overall perspective is that these techniques have the potential to bring robust model comparison techniques based on BF closer to everyday hydrological modelling practice. Aside from the algorithmic developments in this paper, a necessary technological requirement would be the (re-)implementation of hydrological models in a differentiable programming language, e.g. JAX, PyTorch or TensorFlow, rather than in a traditional language such as C, Fortran or Python. While using modern differentiable programming techniques is commonplace for model developers working with machine-learning type approaches, e.g. neural networks, it is less commonly used, but no less applicable, for more traditional hydrological modelling approaches like the ODE-based HBV-like system we consider here. We hope our results encourage more hydrologists to consider differentiable programming tools for conceptual model implementation given the advantages that derivative-based sampling and optimisation algorithms bring to the table in terms of computational efficiency and improved insight, e.g. model selection.

The rest of the paper is organized as follows. Section 2.2 is about conceptual hydrological models and Bayesian methodology, which includes model formulation, prior and likelihood construction, posterior predictive checks, numerical methods, and algorithms. Section 2.3 contains the results and discussions, while the conclusions are provided in Section 2.4. There is also a list of acronyms.

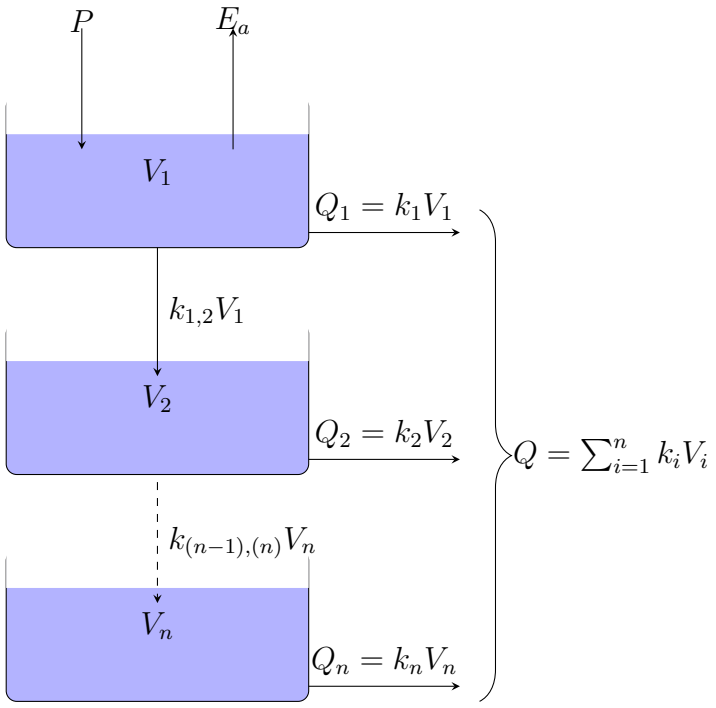


Figure 2.1: Schematic representation of HBV-like ODE model with n-buckets according to the notations in the text. The blue boxes represent the buckets with given state V_1 to V_n. The solid arrows represent mass flows between buckets, into the system or out of the system. The dashed arrow represents the collective mass flow between multiple buckets.

2.2 Methodology

This section describes the model formulation, including prior specification, likelihood construction, algorithms used, and implementation in differentiable software. We leave other modelling aspects, like the type of priors used, for the next section, where we present experiments.

2.2.1 Conceptual models

We develop a set of rainfall-runoff conceptual hydrological models in the framework of continuous dynamical systems that can be written as a system of ODEs of the following form

$$\begin{aligned} V_t &= f(t, V, \theta) \quad \forall t \in (0, \bar{T}], \\ V(t=0) &= \hat{V}, \end{aligned} \tag{2.1}$$

where V are the n system states, $V_t := \frac{\mathrm{d}V}{\mathrm{d}t}$ is the derivative of the state with respect to the time variable t, $\bar{T}$ is the final time, $\hat{V} \in \mathbb{R}^n$ are the initial conditions, f are known functions, and $\theta \in \mathbb{R}^p$ is a vector containing the p model parameters.

For the purpose of the results in this paper, we derive a set of HBV-like models under the principle of conversation of mass. The algorithms developed in this study can be applied to other bucket-type models, e.g. Parajka et al. (2007); Jansen et al. (2021) or those described in the comprehensive MARRMoT rainfall-runoff models toolbox (Trotter et al., 2022). In comparison with the 'standard' HBV model (Bergström, 1976), our model lacks snow and a routing routine and we choose to replace the traditional soil moisture routine with a linear reservoir. A schematic representation of mass flow between the buckets system is given in Fig. 2.1. The system states $\{V_1, \ldots, V_n\}$ $[L^3]$, where L is a generic length unit, represent the volume of water in the i-th bucket and n is the total number of buckets. The system of ODEs for general $n \geq 1$ can be written

$$(V_1)_t = P - E_a - k_1 V_1, \qquad n = 1, \tag{2.2a}$$

$$(V_1)_t = P - E_a - k_1 V_1 - k_{1,2} V_1, \qquad n \geq 2, \tag{2.2b}$$

$$(V_i)_t = k_{(i-1),(i)} V_{i-1} - k_i V_i - k_{(i),(i+1)} V_i, \qquad i = 2, \ldots, n-1, \quad n \geq 3, \tag{2.2c}$$

$$(V_n)_t = k_{(n-1),(n)} V_{n-1} - k_n V_n, \qquad n \geq 2, \tag{2.2d}$$

$$V(t=0) = \hat{V}, \tag{2.2e}$$

$$E_a = \frac{E_p}{V_{\max}} V_1, \tag{2.2f}$$

$$Q = \sum_{i=1}^{n} k_i V_i. \tag{2.2g}$$

The parameters $k_{(i-1),(i)}$ $[T^{-1}]$, $i = 2, \ldots, n$, are the interbucket recession coefficients, where T is a generic time unit. The parameters $k_{(i)}$ $[T^{-1}]$, $i = 1, \ldots, n$, are the outflow recession coefficients. The total outflow Q $[L^3 T^{-1}]$ specified in Eq. (2.2g) is the noiseless quantity y used in the upcoming calibration and model selection procedures. The precipitation P $[L^3 T^{-1}]$ is an *a priori* known function of time. Potential evaporation E_p $[L^3 T^{-1}]$ is a known function of time, whereas actual evaporation E_a $[L^3 T^{-1}]$ is a function of E_p, and $V_{\max}$ $[L^3]$ through Eq. (2.2f), where $V_{\max}$ is the maximum amount of water the soil can store. We remark that the term $E_p/V_{\max}$ in Eq. (2.2f) has units $[L^3 T^{-1}]$ and can therefore be thought of as a dynamic recession coefficient with the dynamic behaviour controlled by the known time-varying potential evapotranspiration function E_p.

The parameter vector $\theta \in \mathbb{R}^p$ associated with the model is then

$$\theta := \{\underbrace{V_{\max}}_{1}, \underbrace{k_1, \ldots, k_n}_{n}, \underbrace{k_{1,2}, \ldots, k_{(n-1),(n)}}_{n-1}, \underbrace{\hat{V}_1, \ldots, \hat{V}_n}_{n}\} \tag{2.3}$$

The number of buckets can be varied by adjusting $n \in \mathbb{N}^+$, leading to a set of models $\{M_1, \ldots, M_n\}$ each with n states and $p = 3n$ parameters. Note that for $i > j$ a more complex model M_i contains a superset of the components of a simpler model M_j. Consequently after calibration of both models on a dataset produced by M_j, M_i should be able to reproduce the data as well as M_j, but at the cost of higher model complexity. This construction will be used in the results to show that the BF does penalise the complex model M_i, leading to the selection of M_j, the expected result.

2.2.2 Bayesian methodology

We briefly restate the Bayes theorem in order to set our notation. If y is the data and θ the parameter vector associated with a model M, then Bayes' theorem in Eq. (2.4) defines the posterior probability of θ as

$$\underbrace{\pi(\theta|y,M)}_{\text{posterior}} = \frac{\overbrace{f(y|\theta,M)}^{\text{likelihood}}\overbrace{\pi(\theta|M)}^{\text{prior}}}{\underbrace{p(y|M)}_{\text{marginal (averaged) likelihood}}} = \frac{f(y|\theta,M)\pi(\theta|M)}{\int f(y|\theta,M)\pi(\theta|M)d\theta}. \tag{2.4}$$

The prior is a probability distribution of a parameter before data is considered. It can incorporate expert knowledge, historical results or any belief about the model parameters. The likelihood tells us how likely various parameter values could have generated the observed data. The denominator in Bayes' theorem

$$p(y|M) = \int \overbrace{f(y|\theta,M)}^{\text{likelihood}}\overbrace{\pi(\theta|M)}^{\text{prior}}\, \mathrm{d}\theta, \tag{2.5}$$

is called the marginal likelihood. The marginal likelihood tells us how likely the model supports the data. The distribution of the parameters given the data is known as the posterior and is proportional to the product of the likelihood and the prior. In the Bayesian paradigm, all inference is based on the posterior.

Likelihood construction

In this section, we drop the explicit index on the model for notational convenience. We define a solution operator $G_{\text{sol}} : \mathbb{R}^{3n} \to X$ that maps a parameter vector θ_j to the total outflow function Q. Concretely, this solution operator is calculated by numerically solving Eqs. (2.2a) to (2.2g). We then define the observation operator $G_{\text{obs}} : X \to \mathbb{R}^q$ which evaluates the solution $Q \in X$ at a set of q points in time $\{t_1, \ldots, t_q\}$.

We assume the following standard Gaussian white noise model for the observed data: $y = G_{\text{obs}}G_{\text{sol}}(\theta) + \eta$ where $\eta \sim \text{MVN}(0, \sigma^2 I_q)$ with MVN a multivariate normal distribution with mean $0 \in \mathbb{R}^q$ and covariance $\sigma^2 I_q \in \mathbb{R}^{q\times q}$, with $\sigma^2 \in \mathbb{R}$ the variance of the measurement noise and I_q the usual q-dimensional identity matrix. Let $G := G_{\text{obs}}G_{\text{sol}} : \mathbb{R}^{3n} \to \mathbb{R}^q$. By standard arguments it can be shown that $y|\theta \sim \text{MVN}(G(\theta), \sigma^2 I_q)$ resulting in the likelihood $f(y|\theta, M)$ in Eq. (2.4) being fully defined. For brevity, we leave precise prior specifications to the results in Section 2.3.

We remark that according to (Cheng et al., 2014) our choice of a likelihood function with Gaussian white noise is equivalent to using the well-known Nash Sutcliffe efficiency (NSE) as a metric. However, other popular metrics such as Kling Gupta efficiency (KGE) cannot be linked explicitly with a likelihood function, and consequently cannot be used within a formal Bayesian analysis. A recent work (Y. Liu et al., 2022) proposes an adaptation of the KGE idea using a Gamma distribution which can be used as an *informal* likelihood function within a Bayesian analysis, but we do not explore this option further here. An alternative option which bypasses the need for an explicit likelihood function is Approximate Bayesian computation (ABC) could be an appropriate alternative when an appropriate explicit metric or likelihood function are unavailable see e.g. (Nott et al., 2012; S. Liu et al., 2023).

Model comparison

The marginal likelihood is also called the normalizing constant (M.-H. Chen et al., 2000; Gelman & Meng, 1998), prior predictive density, evidence (MacKay, 2003) or integrated likelihood (Lenk & DeSarbo, 2000; Gneiting & Raftery, 2007). This quantity is essential to the definition of the Bayes factor. Indeed, the Bayes factor for two competing models, M_i and M_j with $i \neq j$ is the ratio of their marginal likelihoods

$$\text{BF}_{ij} = \frac{p(y|M_i)}{p(y|M_j)} = \frac{\int f(y|\theta_i, M_i)\pi(\theta_i|M_i)d\theta_i}{\int f(y|\theta_j, M_j)\pi(\theta_j|M_j)d\theta_j}. \tag{2.6}$$

Since BF is a ratio, a value greater than one means that M_i should be preferred to M_j, and vice-versa for a value smaller than one. Kass & Raftery (1995) proposed a measure of the strength of evidence (Table 2.1) that we will use throughout this paper to interpret the Bayes factors.

An appealing feature of the BF is its consistency in the limit of a high number of observations. Proofs of consistency for non-nested models are in Casella et al. (2009). For other cases, including nonparametric models, a review and detailed study of consistency can be found in Chib & Kuffner (2016). Also, information theoretic model selection approaches usually require an explicit penalty for the number of model parameters (model complexity). In contrast, the BF implicitly penalises the complexity of the model. That is we do not need to assign a penalty for model complexity since it is already accounted for in the marginal likelihood and hence the BF.

Table 2.1: Interpretation of the Bayes factor (Kass & Raftery, 1995)

$\log_{10}$ BF$_{ij}$	BF$_{ij}$	Evidence in favour of model 1
0 to 1/2	1 to 3.2	Not worth more than a bare mention
1/2 to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
>2	>100	Decisive

Posterior predictive checks

Model selection does not reveal discrepancies between the predictions from the chosen model and observed data. Hence posterior predictive checks (PPCs) are also necessary to see if the selected model can replicate the observed data (Gelman, Meng, & Stern, 1996). PPCs can be graphical or formal. Graphical PPCs consist in making plots of predictions from the chosen model and the observed data to reveal discrepancies. Formal PPC entails calculating a posterior predictive p-value (PPP). The concept of posterior predictive checking was introduced by Rubin (1984) and later generalised by Gelman, Meng, & Stern (1996) under the name PPP where a discrepancy measure can depend on the model parameters. PPCs are the Bayesian equivalent of frequentist goodness-of-fit tests, with the difference that the PPP can be based on any discrepancy measure, not just a statistic.

To compute the PPP, the chosen discrepancy measure is calculated based on replicated data y^{rep}, drawn from the predictive distribution $\pi(y^{\text{rep}}|y_{\text{obs}}) = \int f(y^{\text{rep}}|\theta)\pi(\theta|y_{\text{obs}})d\theta$, and compared with that based on observed data. In mathematical terms, we wish to approximate

the theoretical probability

$$\text{ppp}(y_{\text{obs}}) = Pr\big[D(y^{(\text{rep})}, \theta) \geq D(y_{\text{obs}}, \theta)|y_{\text{obs}}\big]. \tag{2.7}$$

This quantity can be estimated as

$$\text{ppp}(y_{\text{obs}}) = \frac{1}{B}\sum_{i=1}^{B} I\big[D(y_i^{\text{rep}}, \theta_i) \geq D(y_{\text{obs}}, \theta_i)\big] \tag{2.8}$$

where $I[A]$ stands for the indicator function which takes the value 1 if A occurs and 0 otherwise, y_{obs} is the observed dataset, y_i^{rep} is a replicated dataset from the posterior predictive distribution, B is the number of replicated datasets, while θ_i is a single draw from the posterior distribution.

Unlike the frequentist p-value, the interpretation of the PPP is not straightforward since it does not follow a uniform distribution but is concentrated around 0.5 (Meng, 1994). When the p-value has a uniform distribution, the type I error can be controlled. For the frequentist p-value, the probability of falsely rejecting a null hypothesis, which is referred to as a type I error rate, can be set to a fixed value. Typically, this value is prespecified at 0.05 or 0.01. On the contrary, it is difficult to fix the type I error rate for the PPP. Hence, we might fail to reject poor models for a given PPP at a chosen type one error (Gelman, 2013; Hjort et al., 2006). For this reason, we computed the prior calibrated posterior predictive p-value (PCPPP) introduced by Hjort et al. (2006) that has a uniform distribution and the same interpretation as a classical p-value. For more on the Type I error and the distribution of the p-value, refer to Hung et al. (1997) and for Bayesian p-values, see J. L. Zhang (2014). To calculate the PCPPP, a PPP based on data from the prior predictive distribution $\pi(y_{\text{prior}}) = \int f(y^{\text{rep}}|\theta)\pi(\theta)d\theta$ is computed and compared with a PPP based on replicated data from the posterior predictive distribution

$$\text{pcppp}(y_{\text{obs}}) = \frac{1}{B}\sum_{i=1}^{B} I\big[\text{ppp}(y_{\text{prior}_i}^{\text{rep}}) \leq \text{ppp}(y_{\text{obs}})\big],$$

where $\text{ppp}(y_{obs})$ is obtained by Eq. (2.8) and $\text{ppp}(y_{\text{prior}_i}^{\text{rep}})$ can be in a similar way. From this equation, it becomes visible that the PCPPP can also reveal prior data conflicts. A PCPPP greater than a prespecified type I error, say 0.05, means that the prior distribution and model support the data at the level 0.05. The PPP can as well be calibrated based on posterior samples (Hjort et al., 2006; Wang & Xu, 2021).

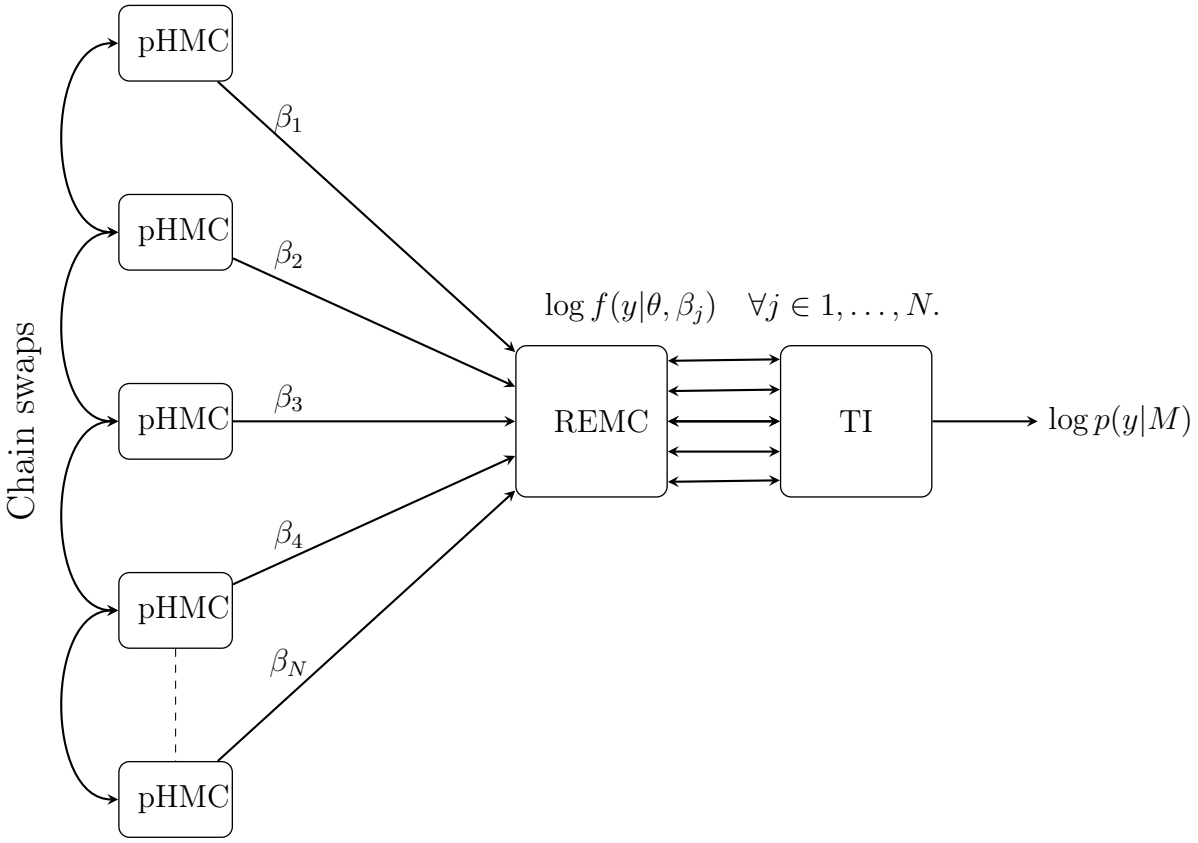


Figure 2.2: Overall schematic of the REpHMC+TI algorithm for estimating the marginal likelihood for a given model M. Working from left to right, N pHMC samplers are run at different values of the inverse temperature parameter $\{\beta_1, \beta_2, \ldots, \beta_N\}$ with $0 \leq \beta_j \leq 1, j = 1, \ldots, N$, to simulate from the power posterior $\log f(y; \theta_i, \beta_j)$. The REMC algorithm is responsible for swapping the state between adjacent chains according to the Metropolis-Hastings criteria. Finally, the TI methodology is used to calculate an estimate of the marginal likelihood $\log p(y|M)$. Note that in terms of setup, information flows from right to left, i.e. the discretisation of the TI integral is responsible for setting the number N and values of inverse temperatures $\beta_1, \ldots, \beta_N$.

2.2.3 Numerical methods

In this section we discuss the proposed new numerical method Replica exchange Hamiltonian Monte Carlo (REHMC) + TI that we employ to simultaneously draw posterior samples and estimate the marginal likelihood. We recommend the reader refer to Fig. 2.2 and its caption for a high-level overview of the approach before continuing to the detailed descriptions below.

Thermodynamic integration

Thermodynamic integration (TI) has its origins in theoretical physics, where it is used to calculate free energy differences between systems (Torrie & Valleau, 1977) before appearing in the statistical literature as path sampling (Gelman & Meng, 1998), a method for calculating marginal likelihoods. TI converts a high-dimensional integral into a one-dimensional integration problem over a unit interval.

To derive the TI estimate of the marginal likelihood $p(y)$, we first raise the likelihood to the power $0 \leq \beta \leq 1$ to form the power posterior (Friel & Pettitt, 2008)

$$\pi_{\text{power}}(\theta|y,\beta) = \frac{\big[f(y|\theta)\big]^{\beta}\pi(\theta)}{p(y|\beta)}, \tag{2.9}$$

with

$$p(y|\beta) = \int \big[f(y|\theta)\big]^{\beta}\pi(\theta)\ \mathrm{d}\theta. \tag{2.10}$$

When $\beta = 0$, the power posterior is the same as the prior distribution. When $\beta = 1$, we have the standard posterior distribution. This makes a continuous path between the prior and the posterior distributions.

Taking the logarithm on both sides of Eq. (2.10) and using the chain rule, a differentiation with respect to β yields

$$\begin{aligned}
\frac{\partial}{\partial\beta}\log p(y|\beta) &= \frac{1}{p(y|\beta)}\frac{\partial}{\partial\beta}p(y|\beta) \\
&= \frac{1}{p(y|\beta)}\int \frac{\partial}{\partial\beta}\big[f(y|\theta)\big]^{\beta}\pi(\theta)\ \mathrm{d}\theta \\
&= \frac{1}{p(y|\beta)}\int \big[f(y|\theta)\big]^{\beta}\log f(y|\theta)\pi(\theta)\ d\theta \\
&= \int \frac{\big[f(y|\theta)\big]^{\beta}\pi(\theta)}{p(y|\beta)}\log f(y|\theta)\ d\theta \\
&= \mathbb{E}_{p(\theta|y,\beta)}[\log f(y|\theta)],
\end{aligned} \tag{2.11}$$

where $\mathbb{E}_{p(\theta|y,\beta)}$ is the expectation with respect to the power posterior. Integrating both sides of equation (2.11) with respect to β gives the log of the marginal likelihood of interest $p(y)$ in terms of an integral on β

$$\log p(y) = \int_0^1 \mathbb{E}_{p(\theta|y,\beta)}[\log f(y|\theta)]\ d\beta, \tag{2.12}$$

This manipulation allows us to find a way to approximate the value of $p(y)$. Computationally, posterior samples are drawn for each value of β. The values are then evaluated in the log-likelihood, and the mean for each value of β is obtained. The integral (2.12) on β can be estimated using the trapezoidal rule as follows:

$$\log p(y) = \sum_{j=1}^{N} \frac{(\beta_j - \beta_{j-1})}{2} \left[\mathbb{E}_{p(\theta|y,\beta_j)} \log f(y|\theta) + \mathbb{E}_{p(\theta|y,\beta_{j-1})} \log f(y|\theta) \right].$$

The Monte Carlo estimate of the expectations can then be obtained by

$$\log p(y) \approx \sum_{j=1}^{N} \frac{(\beta_j - \beta_{j-1})}{2} \left[\frac{1}{S} \sum_{i=1}^{S} \log f(y|\theta_i, \beta_j) + \frac{1}{S} \sum_{i=1}^{S} \log f(y|\theta_i, \beta_{j-1}) \right], \tag{2.13}$$

where $j = 1, \ldots, N$ is the index for the β values and S is the number of posterior samples for each β value. The accuracy of the TI estimate depends on the integration rule on β, i.e. the number of β values and the spacing of the values, and the convergence of the Markov Chain Monte Carlo (MCMC). The most commonly employed path is a geometric path (Calderhead & Girolami, 2009)

$$\beta_j = \left(\frac{j}{N} \right)^5, \quad j = 1, \ldots, N. \tag{2.14}$$

The number of β_j values can be adaptively chosen as a tradeoff between model convergence and computational efficiency, for instance, see Vousden et al. (2016). The complete TI algorithm is presented in Algorithm 1.

Algorithm 1 Thermodynamic integration (TI)

Input: β $\{\beta = \{1, \ldots, 0\}$: schedule of inverse temperatures based on trapezoidal rule of size N, S is the number of samples per replica.$\}$

Output: Log marginal likelihood ($\log p(y)$).

1: REpHMC(β) {Run the a single step of the REpHMC algorithm S times, see section 2.2.3.}

2: Estimate $\log p(y)$ by the definition of the quadrature rule, e.g. trapezoidal rule

$$\log p(y) \approx \sum_{j=1}^{N} \frac{(\beta_j - \beta_{j-1})}{2} \left[\frac{1}{S} \sum_{i=1}^{S} \log f(y|\theta_i, \beta_j) + \frac{1}{S} \sum_{i=1}^{S} \log f(y|\theta_i, \beta_{j-1}) \right].$$

Replica exchange Monte Carlo

The REMC algorithm was introduced by Swendsen & Wang (1986). Geyer (1991) presented a similar formulation to the statistical community under the name Metropolis-coupled MCMC. REMC is a generic algorithm in that it can be combined with other algorithms. Miasojedow et al. (2013) combined REMC with random walk Metropolis (RWM). RWM is a gradient-free algorithm in that it generates posterior samples from the target distribution by randomly sampling from a proposal distribution. We combine REMC with HMC, which gives the new algorithm REHMC explained in the rest of this section. When REMC is combined with pHMC, we get the REpHMC. The REpHMC gives a higher effective sample size than REHMC. The effective sample size is the number of independent samples with the same amount of information as correlated samples. Each sample in a Markov chain is correlated to the preceding sample, so the samples have less information than independent samples. The effective sample size takes into account this autocorrelation. The main idea of REMC is that an ensemble of power posterior chains known as replicas run in parallel. The likelihood of these chains is raised to values from zero to one. These values are called inverse temperatures. Each replica performs a Metropolis update to get the next value at each iteration. The replica pairs are randomly selected, and an attempt is made to swap the current values of the replica pairs. A swap is accepted or rejected according to the Metropolis-Hastings algorithm. The swapping accelerates convergence to the target distribution, avoids chains becoming trapped in topologically isolated areas of the parameter space, and improves the mixing of the chains. REMC is also known as parallel tempering (Hansmann, 1997; Earl & Deem, 2005). When the method has an iterated importance sampling step, it is known as population Monte Carlo (PMC) (Iba, 2000; Cappé et al., 2004). However, the term PMC has also been used for methods without an importance sampling step (Calderhead & Girolami, 2009; Friel & Pettitt, 2008; Mingas & Bouganis, 2016).

The REpHMC is summarised in Algorithm 2. We emphasise that the samples of the replica with $\beta = 1$ are used to estimate the posterior parameters, while the entire ensemble is used as input within TI to calculate the marginal likelihood.

Like any sampling method, the REpHMC's convergence should be assessed. We used both trace plots and formal diagnostic tests to check for convergence of the Markov chain since there is no universal robust test for convergence (Cowles & Carlin, 1996). The most widely used method to assess the convergence of Markov chains is the potential scale reduction factor $\hat{R}$, developed by Gelman & Rubin (1992) and extended by Brooks & Gelman (1998). Recently,

Algorithm 2 Single step of Replica Exchange preconditioned Hamiltonian Monte Carlo (REpHMC)

Input: L, ϵ, θ^t, β {L: number of leapfrog steps, ϵ: leapfrog stepsize, $\theta^t = \{\theta_1^t, \ldots, \theta_N^t\}$: initial values for each β, $\beta = \{\beta_1, \ldots, \beta_N\}$: schedule of N inverse temperatures}
Output: $(\theta_1^{t+1}, \ldots, \theta_N^{t+1})$ {Posterior samples for each β}.

for $i = 1$ **to** N **do**
  $\theta_i^{t+1} \leftarrow \text{pHMC}(L, \epsilon, \theta_i^t)$ {Run single step of pHMC algorithm on each replica}
end for
for $i = 1$ **to** $N - 1$ **do**
  $j \leftarrow i + 1$ {Select adjacent chain}
  $\alpha \leftarrow \min\left(1, \frac{\pi_i(\theta_j^{t+1})\pi_j(\theta_i^{t+1})}{\pi_i(\theta_i^{t+1})\pi_j(\theta_j^{t+1})}\right)$ {where e.g. $\pi_i(\cdot)$ is the power posterior associated with temperature β_i}.
  $u \sim U(0, 1)$
  if $u \leq \alpha$ **then**
    $(\theta_i^{t+1}, \theta_j^{t+1}) \leftarrow (\theta_j^{t+1}, \theta_i^{t+1})$
  else
    $(\theta_i^{t+1}, \theta_j^{t+1}) \leftarrow (\theta_i^{t+1}, \theta_j^{t+1})$
  end if
end for

an improved factor $\hat{R}$ was proposed by Vehtari et al. (2021). For $\hat{R}$ to be a valid statistic, the chains must be independent of each other. In REHMC, the chains are not independent due to swapping. Therefore, we used methods that require one chain or replica per temperature, namely the Geweke diagnostic (Geweke, 1991) and the IAT (Geyer, 1992; Kendall et al., 2005). For the sake of brevity, we do not explain these concepts here but instead refer the reader to the respective papers.

Hamiltonian Monte Carlo

HMC is a gradient-based technique used to sample from a continuous probability density (Duane et al., 1987). HMC scales better in high dimensions than gradient-free samplers, such as nested sampling, due to the inclusion of derivative information (Ashton et al., 2022). Therefore, many applications combine HMC and gradient-free samplers. For example, Elsheikh et al. (2014) has combined HMC and nested sampling. HMC is based on the Hamiltonian, which describes a particle's position and momentum at any time. New positions are known by solving Hamilton's equations of motion for position and momentum. In Bayesian inference, the Hamiltonian $H(\theta, \rho)$ in Eq. (2.15) describes the evolution of a d dimensional vector (θ) of parameters and a corresponding d dimensional vector of auxiliary momentum (ρ) variables at any time, t.

$$\begin{aligned} H(\theta, \rho) &= -\log f(y|\theta)\pi(\theta) + \frac{1}{2}\rho^T M \rho \\ &= U(\theta) + K(\rho) \end{aligned} \tag{2.15}$$

In Eq. (2.15), M is the positive definite mass matrix. $U(\theta)$ is the desired posterior known as potential energy, and $K(\rho)$ is the kinetic energy that is a function of momentum. To sample from the Hamiltonian, we take the partial derivatives, which give Hamilton's equations of motion

$$\frac{d\theta}{dt} = \frac{\partial H}{\partial \rho} = \frac{\partial K}{\partial \rho} \tag{2.16a}$$

$$\frac{d\rho}{dt} = -\frac{\partial H}{\partial \theta} = -\frac{\partial U}{\partial \theta} \tag{2.16b}$$

We now have a system of ODEs (Eqs. (2.16a) to (2.16b)). The leapfrog method (Duane et al., 1987; Radford M. Neal, 2011) is used to solve the Eqs. (2.16a) to (2.16b) and propose new values for the parameters. The accuracy of the leapfrog method depends on the discretisation step ϵ.

Each HMC iteration consists of two steps (Radford M. Neal, 2011). In the first step, the momentum values for each parameter are sampled from a Gaussian distribution independent of the current θ values, $\rho^* \sim MVN(0, M)$.Then using the current parameter and momentum values, (θ^t, ρ^t), the Hamiltonian is simulated using an appropriate time stepping method such as the leapfrog method (Betancourt, 2017). At the end of Hamiltonian dynamics, the momentum values are negated, and the new parameter values (θ^*, ρ^*) are accepted or rejected using the Metropolis-Hastings criterion with acceptance probability α where

$$\alpha = \min\left[1, \exp\left(-U(\theta^*) + U(\theta^t) - K(\rho^*) + K(\rho^t)\right)\right]. \tag{2.17}$$

The HMC is summarised in Algorithm 3. The mixing of the HMC chain depends on the number of leapfrog steps L and the step size ϵ. L and ϵ can be automatically tuned during the warm-up phase of the algorithm (Hoffman & Gelman, 2014). The warm-up phase is the period during which posterior samples are discarded and is also called burn-in. In this work, ϵ was automatically tuned by the dual averaging algorithm while L was manually tuned. Dual averaging automatically adjusts ϵ during the warm-up of the HMC algorithm until a specific acceptance rate is achieved. We used an acceptance rate of 0.75, which is higher than the optimal acceptance rate of RWM based algorithms. This is the mean of various reported values and the default in TensorFlow probability. To increase the sampling efficiency of HMC, we have to reduce the correlation of the parameters, especially for ODE models. This is achieved

by introducing a preconditioned matrix, M and hence the name pHMC. This leads to even faster convergence and higher effective sample sizes for each parameter (Girolami & Calderhead, 2011). In practice, the preconditioned matrix is the inverse of the covariance matrix of the target posterior. In contrast to HMC, where the momentum is sampled from a normal distribution, for pHMC, the momentum values are sampled from a multivariate Gaussian distribution with a covariance matrix as the preconditioned matrix, $\rho \sim \text{MVN}(0, M)$. The covariance matrix controls the correlation of the parameters. The rest of the algorithm for pHMC works as for HMC.

Algorithm 3 Single step of preconditioned Hamiltonian Monte Carlo (pHMC), Notation following Radford M. Neal (2011)

Input: L, ϵ, θ^t {L: number of leapfrog steps, ϵ: leapfrog step size, θ^t: initial value.}
Output: θ^{t+1}

$\rho^* \sim MVN(0, M)$ {Sample momentum values, M is the mass matrix}
$\theta^* \leftarrow \theta^t$
for $i = 1$ **to** L **do**
    $(\theta^\epsilon, \rho^\epsilon) \leftarrow \text{Leapfrog}(\theta, \rho, \epsilon)$
end for
$\rho^* \leftarrow -\rho^*$
$\alpha \leftarrow \min\left(1, \exp\left(-U(\theta^*) + U(\theta^t) - K(\rho^*) + K(\rho^t)\right)\right)$
$u \sim U(0, 1)$
if $u \leq \alpha$ **then**
    $\theta^{t+1} \leftarrow \theta^*$
else
    $\theta^{t+1} \leftarrow \theta^t$
end if

function Leapfrog(θ, ρ, ϵ) {Solves the equations to propose new values}
    $\rho^{\epsilon/2} \leftarrow \rho - \frac{\epsilon}{2}\frac{\partial U}{\partial \theta}(\theta)$
    $\theta^\epsilon \leftarrow \theta + \epsilon M^{-1}\rho^{\epsilon/2}$
    $\rho^\epsilon \leftarrow \rho^{\epsilon/2} - \frac{\epsilon}{2}\frac{\partial U}{\partial \theta}(\theta^\epsilon)$
    return $(\theta^\epsilon, \rho^\epsilon)$

2.2.4 Implementation aspects

In this section, we outline some of the more non-standard aspects of implementing the proposed methodology in the probabilistic programming language (PPL) TFP. Probabilistic programming (PP) is a methodology for performing computational statistical modelling in which all elements of the Bayesian joint posterior $\pi(\theta|y, M)$ are specified in a PPL. Popular PPLs include Stan (Carpenter et al., 2017), PyMC3 (Salvatier et al., 2016) and TFP (Dillon et al., 2017).

Once specified in a PPL, the subsequent Bayesian parameter inference problem can then be handled semi-automatically. We refer the reader to the Code and Data availability statement for the full implementation and simply remark that the joint posterior for our problem can be defined in around 70 lines of TFP/JAX code.

We choose to use TFP in this study. From our experience, TFP is the most flexible and extensible PPL in terms of allowing advanced model specification and the ability to break out of the high-level interface and perform low-level operations. However, this flexibility comes at the cost of a steeper learning curve, particularly TFP's complex batch and event shape semantics (Dillon et al., 2017). We note that despite TensorFlow in the name, TFP is backend-agnostic and can run on top of various differentiable programming languages. We choose to run TFP on top of JAX, instead of the default choice of TensorFlow. Anecdotally, our experience is that TFP on JAX has better runtime performance and is more robust than TFP on TensorFlow, particularly when working with ODE-based models. We use JAX with the CPU backend and double precision floating point representation, although in principle the GPU backend could also be used. TFP already includes an implementation of the HMC and REMC algorithms, the output of which can be used with TI for computing the marginal likelihood.

JAX can automatically perform arbitrarily composable forward and backward mode automatic differentiation of nearly arbitrary computer programs. This is used to automatically differentiate the TFP specification of the negative log posterior $U(\theta)$ with respect to the model parameters θ for use within the HMC algorithm. As this approach is now standard, we refer the reader to Margossian (2019) for a detailed review.

For the automatic differentiation of the ODE model, we use the continuous adjoint approach. This approach is also called continuous backpropogation in the Neural ODE literature, see e.g. Kelly et al. (2020) and Höge et al. (2022) for an application in hydrology. We follow the presentation in (Kidger, 2021) where a new set of adjoint ODEs is from the original continuous ODE system. This adjoint system is then discretised (backwards in time) using the same ODE solver as for Eq. (2.1), an explicit adaptive Dormund-Prince ODE integrator that is already included in JAX. It is worth remarking that while the continuous adjoint system is still derived automatically within JAX, the result is distinctly different to backwards differentiation through the steps of the forward ODE solver at the programmatic level. For more details, we refer the reader to Kidger (2021) for a discussion of the different methods for automatically differentiating ODE systems and their relative tradeoffs.

Let V be the solution to Eq. (2.1). In the simplest case let $J = J(V(T))$ be some scalar

function of the terminal solution value $V(T)$ (the approach extends straightforwardly to other functionals). Setting $\frac{\mathrm{d}J}{\mathrm{d}V} = a_V(t)$ and $\frac{\mathrm{d}J}{\mathrm{d}\theta} = a_\theta(0)$ where $a_V : [0,T] \to \mathbb{R}^n$ and $a_\theta : [0,T] \to \mathbb{R}^p$ are the solutions to the following adjoint ODE system

$$(a_V)_t = -a_V(t)^T \frac{\partial f}{\partial V}(t, V, \theta), \qquad a_V(T) = \frac{\mathrm{d}J}{\mathrm{d}V(T)}, \tag{2.18a}$$

$$(a_\theta)_t = -a_V(t)^T \frac{\partial f}{\partial \theta}(t, V, \theta), \qquad a_\theta(T) = 0. \tag{2.18b}$$

Note that the adjoint system requires the forward solution to have already been computed and that the adjoint system runs backwards in time, i.e. evolving from known states $a_V(T)$ and $a_\theta(T)$ at terminal time $t = T$ to the starting time $t = 0$. Once $a_\theta(0)$ has been computed, the required gradient of the functional $\frac{\mathrm{d}J}{\mathrm{d}\theta} = a_\theta(0)$ can be computed straightforwardly. This continuous adjoint ODE approach can be arbitrarily composed with JAX's programme level automatic differentiation capabilities, meaning that it is possible to add non-ODE based components (smoothers etc.) to the model and use our framework for computing marginal likelihoods.

2.3 Results and discussion

The purpose of this section is to test the accuracy of REpHMC in calculating the BF by employing it to solve benchmark problems with complex distributions but well known log marginal likelihoods and thus the BF. We illustrate that the BF can distinguish between models with an equally good fit by calculating the BF of synthetic discharge data for three different models, among which is the data generating model. We repeat the experiment using another data generating model. Finally, the BF is applied to the real-world discharge data.

2.3.1 Gaussian shells example

This section aims to show that the the proposed methodology accurately estimates the marginal likelihood of a synthetic example. In addition, it illustrates the effectiveness of REpHMC in sampling multimodal distributions. The benchmark example is the Gaussian shells (Feroz et al., 2009; Allanach & Lester, 2008). This example has two wholly separated Gaussian shells, making it difficult to sample from. This example has been used to test various techniques for calculating the marginal likelihood (Thijssen et al., 2016; Henderson & Goggans, 2019). The

Gaussian shell likelihood is given as

$$\ell(\theta) = \frac{1}{\sqrt{2\pi w_1^2}}\exp\left[-\frac{(||\theta_1 - c_1|| - r_1)^2}{2w_1^2}\right] + \frac{1}{\sqrt{2\pi w_2^2}}\exp\left[-\frac{(||\theta_2 - c_2|| - r_2)^2}{2w_2^2}\right]. \quad (2.19)$$

The unknown parameters are $\theta = (\theta_1, \theta_2)$, while the marginalised fixed parameters are r_1, r_2, w_1, w_2, c_1 and c_2. The first shell has a radius of r_1 and the second shell r_2. The first shell is centred at c_1 while the second is centred at c_2. The variance (width) of the first shell is w_1, and that of shell two is w_2. We assign uniform priors to θ_1 and θ_2 in the range -6 to 6 and the marginalised parameters are set to $w_1 = w_2 = 0.1, r_1 = r_2 = 2, c_1 = 3.5, c_2 = -3.5$. We used 26 temperature schedules, since this is a difficult sampling problem to obtain fast mixing due to the two regions of probability mass. Convergence of the number of temperatures was checked after the convergence of the posterior samples. The log marginal likelihood is stable after using 22 temperatures. From this point, there is very little variation in the log marginal likelihood, as shown in Fig. 2.3. The plot shows that the log marginal likelihood is constant from 10 to 11 temperatures. Although 10 temperatures are commonly used, this would have underestimated the actual value. To assess convergence, diagnostic plots were made by running the same temperature schedules twice in parallel with two different random initial parameter values, and the results are displayed in Fig. 2.3 where the horizontal red line is the true value.

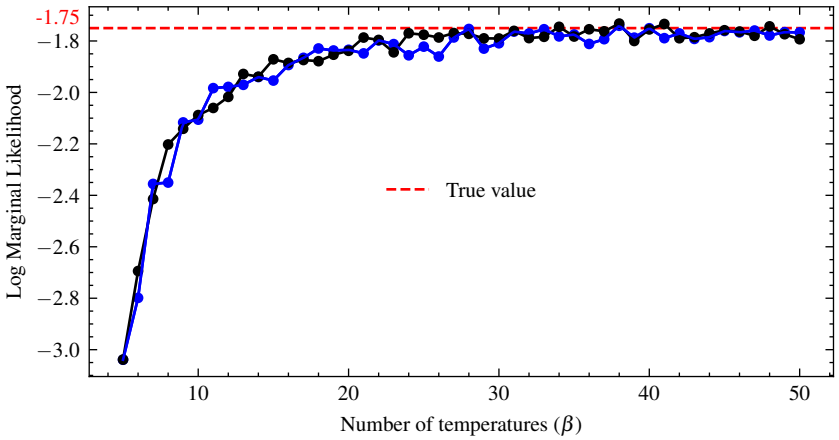


Figure 2.3: Convergence diagnostic plots of the log marginal likelihood for the Gaussian shell in two dimensions. The temperature schedules is run twice in parallel with random initial parameter values. Convergence occurs when the curves plateau.

A plot of the samples for the parameters using various samplers is shown in Fig. 2.4. The plot demonstrates that due to the addition of Replica Exchange the REpHMC method can sample across the the shells, compared to algorithms such as NUTS (Hoffman & Gelman, 2014),

MALA (Xifara et al., 2014) or plain HMC (not shown) which are purely local. The results of the marginal likelihood up to 30 dimensions are shown in Table 2.2 with agreement with the marginal likelihood values reported in the literature (Feroz et al., 2009).

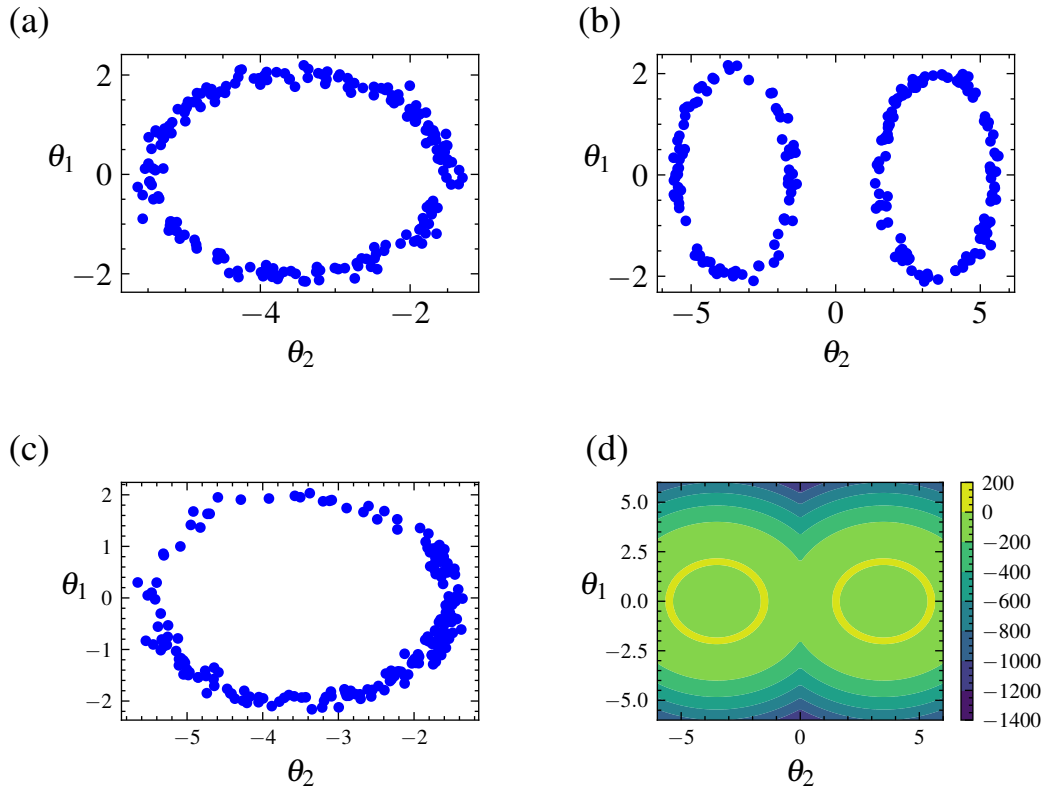


Figure 2.4: Posterior samples for the Gaussian shells example obtained by different algorithms alongside the target distribution. Top left (a) is NUTS, top right (b) is REpHMC, bottom left (c) is MALA and bottom right (d) is the target distribution. Because of the addition of Replica Exchange, REpHMC can sample across the entire distribution space. This is in contrast to the NUTS, MALA and HMC (not shown) samplers which cannot transition across the gap between the two shells.

2.3.2 Synthetic examples

In this section we generate synthetic discharge data by using the observed precipitation and observed potential evapotranspiration as inputs to our models. The following two examples aim to verify the correct implementation and study the behaviour of the methodology to calculate the marginal likelihood. In the first experiment, data y_{obs} is generated from the simplest model, M_2. In the second experiment, M_3 (three buckets model) is the data generating model. For each experiment, the log-marginal likelihood $\log p(y|M_i)$ for $i = 2, 3, 4$ and the respective Bayes factors are calculated. The DIC and WAIC are also calculated for experiments in Section 2.3.2,

Table 2.2: Log marginal likelihood (log $p(y)$) of the Gaussian shell example. The true values are shown, and the estimates are based on thermodynamic integration with samples from REpHMC. The results are shown for up to 30 dimensions.

Dimensions	*Reference log $p(y)$	Estimated log $p(y)$
2	-1.75	-1.75 ± 0.003
5	-5.67	-5.68 ± 0.006
10	-14.59	-14.60 ± 0.006
20	-36.09	-36.12 ± 0.014
30	-60.13	-60.19 ± 0.025

* As reported in Feroz et al. (2009).

Section 2.3.2 and for real-world discharge data in Section 2.3.3.

Experiment one with data generated from the two-buckets model M_2

In the first experiment, synthetic discharge data y_{obs} is generated from the simplest model, M_2 (two buckets model) to see if the BF will select M_2. We set up the priors as in Table 2.3. The synthetic discharge is generated to have similar dynamics as the observed discharge shown in Fig. 2.5. First, we obtain the daily precipitation and evapotranspiration for the Magela Creek catchment in Australia for 1980. The initial time $t = 0$ corresponds to midnight on January 1, 1980, and the final time $T = 366$ days to midnight on December 31, 1980 (1980 had a leap year). It is assumed that the total precipitation and evapotranspiration on a given day is uniformly distributed over the 24 hours from midnight to midnight. This is an acceptable assumption when modelling the dynamics of a catchment on a multiday time scale.

Our analysis focuses on a three-month period in 1980. This period runs from 01-01-1980 to 31-03-1980 when the precipitation frequency is highest, and there are no missing data.

We set up the priors according to the following reasoning:

- The top bucket associated with state V_1 typically represents the fast dynamics of the catchment system, such as surface runoff into rivers. The parameters k_1 and k_{12} are the recession coefficients of the top bucket. They represent the flow rates from the top bucket. Since the parameters have to be positive, we use lognormal priors, the most commonly used distribution for dynamic models.

- The lower buckets represent processes with progressively slower dynamics, such as groundwater storage. The parameters k_i and $k_{(i-1),(i)}$ are the recession coefficients for the n^{th} bucket with $i = 2, 3, \ldots, n$.

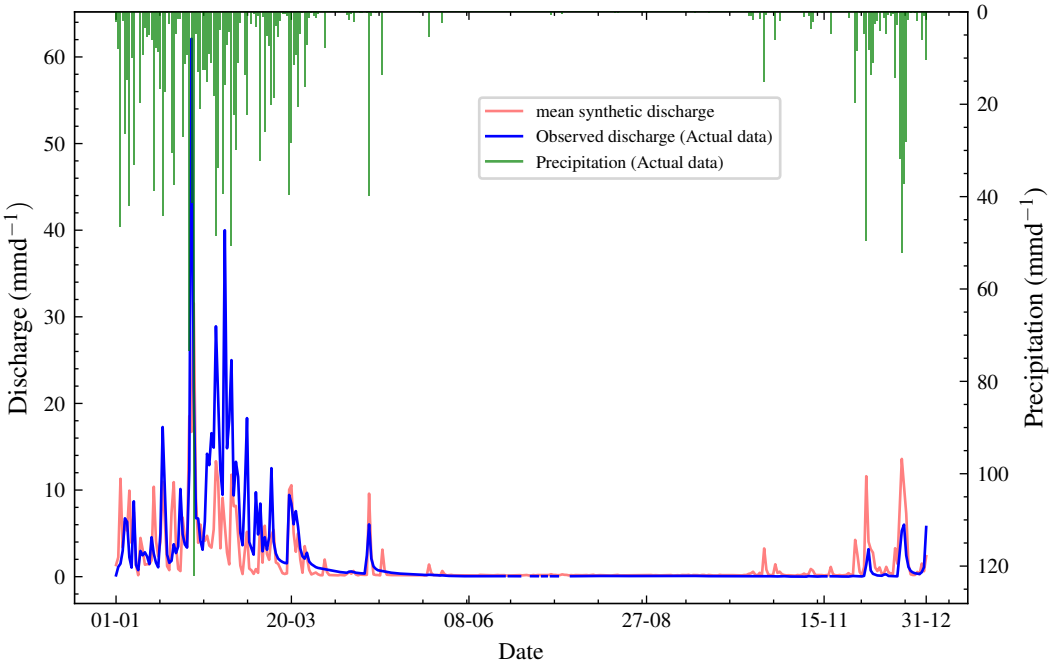


Figure 2.5: Plot of observed discharge, synthetic discharge, and precipitation from 01-01-1980 to 31-12-1980. The observed discharge has missing values, represented by the broken blue line, mostly in the seventh month. Synthetic discharge data generated via the joint posterior (before calibration) shows similar overall trends to the observed discharge.

- The system starts with a nonzero initial condition that mimics the standard procedure of "bootstrapping" the ODE system for a period $T_B < 0$. For real-world data, the initial conditions are not known and must be identified. The initial condition to be identified is $\hat{V}_i$ where $i = 1, 2, \ldots, n$.

The meaning of the parameters and the priors are shown in Table 2.3. We follow a Bayesian workflow and do a prior predictive check. This helps to verify if the priors are reasonable. For the prior predictive check, 50 samples were drawn from the prior and then evaluated in the likelihood. This gave 50 different data sets for the synthetic discharge. The mean synthetic discharge is then obtained, and the 95 % pointwise credible intervals are obtained and shown in Fig. 2.6. The marginal likelihoods for M_2, M_3 and M_4 were calculated and the corresponding Bayes factors were calculated. For each model, fifteen different runs of the marginal likelihood were calculated using REpHMC + TI. This enabled us to get the estimate's standard deviation, which is different from the Monte Carlo standard error.

We perform REpHMC with 10 replicas where the likelihood of a replica is raised to an inverse temperature value according to the schedule in Eq. (2.14). Each replica was run until IAT $< N/50$, where N is the number of posterior samples. The IAT is the number of samples

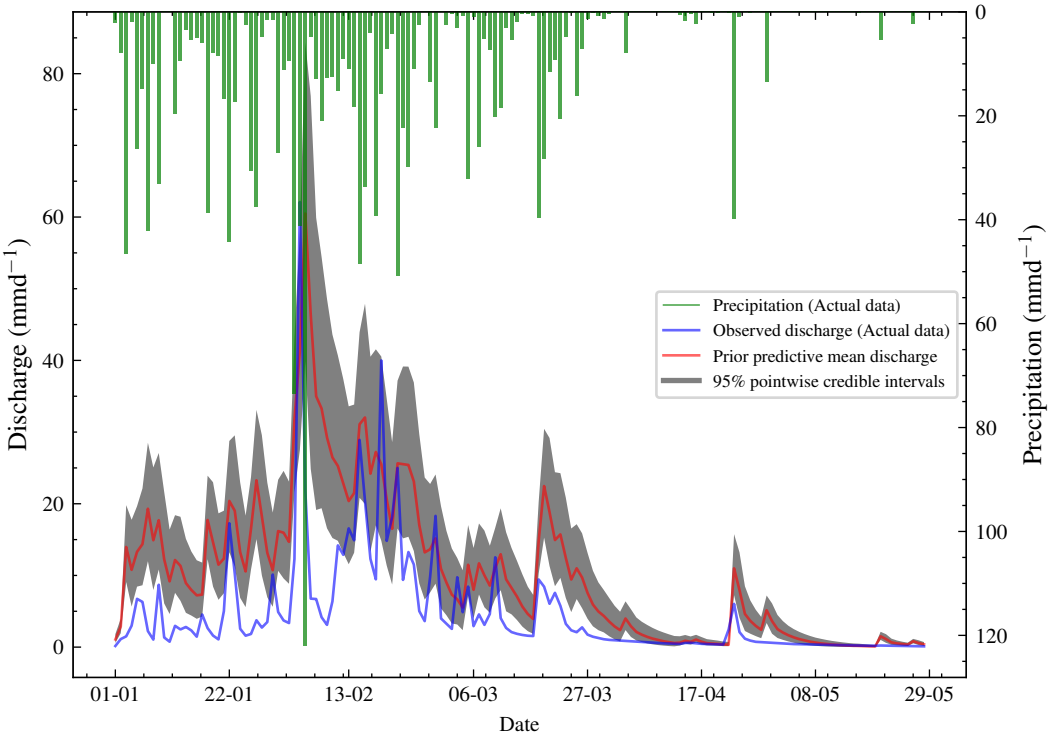


Figure 2.6: Plot of observed discharge, synthetic discharge, and precipitation from 01-01-1980 to 29-05-1980. This period has no missing values and has the highest precipitation frequency and discharge of the year 1980. The synthetic discharge has a similar trend to the observed discharge. The synthetic discharge here is generated using a different set of parameters compared to that in Fig 2.5.

required to obtain an independent sample and a smaller value is preferable. We found that 4000 posterior samples per replica were enough to rule out non-stationarity. We also did a full run with 20000 posterior samples per chain, and we saw no significant change in the results. The p-value for Geweke diagnostics was not significant at 5 % for all parameters and models (p-value > 0.90), indicating there is a high probability that the parameters have converged. The IAT and Geweke diagnostics were performed using the Python package, pymcmcstat (Miles, 2019). The posterior parameter estimates and 95 % credible interval (CI) are in table Table 2.4. For M_2, the true model, the posterior parameters are very close to the true values and are within the 95 % CI. Moreover, the parameters k_1, $\hat{V}_1$, $\hat{V}_2$, $V_{\max}$ and σ^2 are very close to the true values. However, the error term σ^2 is the same for all three models, as all models fit the data well. Therefore, a model selection criterion is needed to discriminate between models. Fifteen marginal likelihoods are calculated in parallel for each model. The mean log marginal likelihood is presented in Table 2.4. We can calculate the log BF of any model compared to another by taking the difference in their log marginal likelihoods. Based on the interpretation

Table 2.3: Description of the parameters and priors. Note that here we have used units more common in the hydrological literature. LN is the lognormal distribution and IG is the inverse Gamma distribution. The IG was chosen because it is easier to sample than other distributions for the prior noise parameter, which must be positive.

Parameter	Unit	Description	Prior
k_1	d^{-1}	Outflow recession coefficient for bucket 1	$\mathrm{LN}(1.0, 0.25)$
k_2	d^{-1}	Outflow recession coefficient for bucket 2	$\mathrm{LN}(0.6, 0.25)$
k_3	d^{-1}	Outflow recession coefficient for bucket 3	$\mathrm{LN}(0.3, 0.25)$
k_4	d^{-1}	Outflow recession coefficient for bucket 4	$\mathrm{LN}(0.1, 0.25)$
k_{12}	d^{-1}	Interbucket recession coefficient 1 to 2	$\mathrm{LN}(0.8, 0.25)$
k_{23}	d^{-1}	Interbucket recession coefficient 2 to 3	$\mathrm{LN}(0.4, 0.25)$
k_{34}	d^{-1}	Interbucket recession coefficient 3 to 4	$\mathrm{LN}(0.1, 0.25)$
$\hat{V}_1$	mm	Initial condition on V_1	$\mathrm{LN}(0.0, 1.0)$
$\hat{V}_2$	mm	Initial condition on V_2	$\mathrm{LN}(0.0, 1.0)$
$\hat{V}_3$	mm	Initial condition on V_3	$\mathrm{LN}(0.0, 1.0)$
$\hat{V}_4$	mm	Initial condition on V_4	$\mathrm{LN}(0.0, 1.0)$
$V_{\max}$	mm	Maximum amount of water the soil can store	$\mathrm{LN}(1.0, 0.25)$
σ^2	$\mathrm{mm}^2\mathrm{d}^{-2}$	Variance of the Gaussian noise model	$\mathrm{IG}(5.0, 0.1)$

of BF in Table 2.1, there is decisive evidence in favour of the data generating model M_2. The distributions of the log marginal likelihood for each model are shown in box plots (Fig. 2.7). In addition, the DIC and WAIC are shown along with those of the marginal likelihood and they also select the data generating model. The DIC is a Bayesian generalisation of information-theoretic based criterion AIC for model selection introduced by Spiegelhalter et al. (2002). The WAIC is based on pointwise out-of-sample predictive accuracy (Vehtari et al., 2017; Watanabe & Opper, 2010) and for large samples equivalent to the leave out one cross-validation (Watanabe & Opper, 2010). For these information-based theoretic methods, a difference of 10 is usually required for a decisive preference of one model over the other ("Information and Likelihood Theory: A Basis for Model Selection and Inference", 2002, p. 70). A difference of up to 7 is considered less support to prefer one model over the other (Spiegelhalter et al., 2002). Model M_2 has the largest median log marginal likelihood, while model M_4 has the lowest. The prior and posterior distributions for model M_2 are in Fig. 2.8. The prior distribution is in blue, while the posterior is in red. The prior range is wide compared to the posterior such that the posterior contours are too small. The posterior marginal densities are also more contracted compared to the prior densities, as seen on the diagonal of the plots. The prior contours show no significant correlation between the parameters. The posterior distributions for this model are shown in Fig. 2.9. The marginal

posterior distributions are on the diagonal. The red dots represent the true parameters. There is also a high correlation between pairs $(k_1, k_2), (k_1, V_{\max}), (k_{1,2}, k_2), (k_{12}, V_{\max}), (k_2, V_{\max})$ and $(\hat{V}_1, \hat{V}_2)$.

Table 2.4: True value, posterior mean with 95 % credible intervals of the parameters, and log marginal likelihood of the models for experiment one. Model M_2 has the highest log marginal likelihood and is the true model. The DIC and WAIC are also shown.

parameter	True value	M_2 (95 % CI)	M_3 (95 % CI)	M_4 (95 % CI)
k_1	1.454	1.454 (1.445, 1.462)	1.438 (1.434, 1.457)	1.089 (1.081, 1.095)
k_2	0.248	0.248 (0.248, 0.248)	0.241 (0.241, 0.250)	0.160 (0.129, 0.174)
k_3	0.000	-	0.248 (0.247, 0.248)	0.241 (0.196, 0.265)
k_4	0.000	-	-	0.208 (0.207, 0.208)
$k_{1,2}$	3.232	3.234 (3.205, 3.263)	3.157 (3.145, 3.256)	1.628 (1.552, 1.670)
$k_{2,3}$	0.000	-	1.619 (0.993, 1.683)	1.102 (0.921, 1.400)
$k_{3,4}$	0.000	-	-	1.861 (1.105, 2.749)
$\hat{V}_1$	1.081	1.067 (1.039, 1.095)	1.067 (1.038, 1.071)	1.246 (1.181, 1.282)
$\hat{V}_2$	0.813	0.894 (0.787, 0.990)	0.490 (0.483, 0.593)	0.599 (0.474, 0.761)
$\hat{V}_3$	0.000	-	0.520 (0.453, 0.525)	0.731 (0.459, 0.827)
$\hat{V}_4$	0.000	-	-	0.576 (0.433, 0.954)
$V_{\max}$	2.520	2.520 (2.502, 2.542)	2.573 (2.507, 2.581)	3.106 (2.999, 3.149)
σ^2	0.014	0.014 (0.011, 0.016)	0.015 (0.015, 0.016)	0.023 (0.022, 0.027)
$\log p(y\|M)$	-	217.968	203.383	154.768
DIC	-	-521.235	-448.980	-449.000
WAIC	-	-514.354	-501.686	-445.233

-: The parameter is not included in the model.
σ^2: error term.
log$p(y|M)$: log marginal likelihood.

We also performed graphical posterior predictive checks. Discharge data was generated from the posterior predictive distribution of each model and plotted. There is no noticeable visual difference in discharge (Fig. 2.10) for all the models since the posterior error estimate is too small for all models. We also calculated PPP for the selected model using autocorrelation as a discrepancy measure. Hence, Eq. (2.8) becomes

$$\text{ppp}(y_{\text{obs}}) = \frac{1}{n}\sum_{i=1}^{n} I\left[(\rho_i^{\text{rep}}, \theta_i) \geq (\rho_{\text{obs}}, \theta_i)\right] \tag{2.20}$$

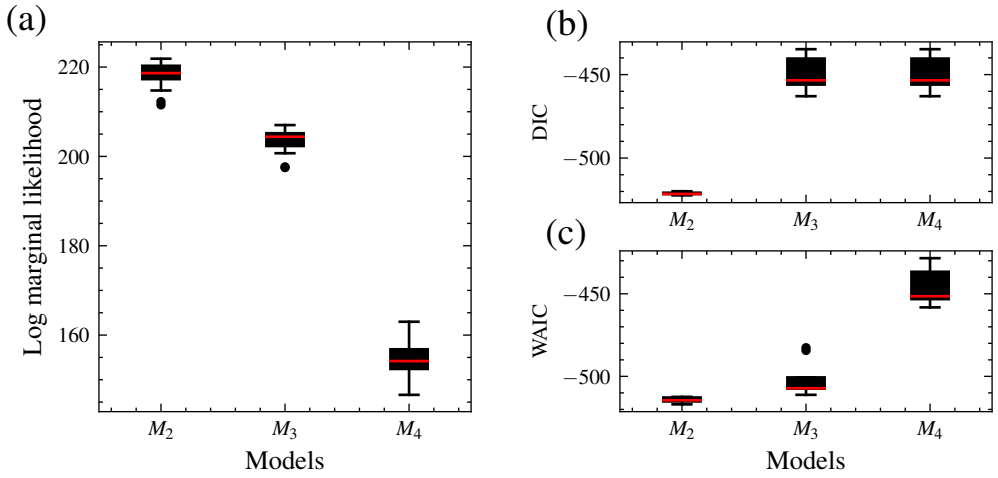


Figure 2.7: Distribution of the log marginal likelihood, DIC and WAIC for 15 different runs each. Distribution of the log marginal likelihood for 15 different runs. The boxplot of the data generating model, M_2, is the highest while M_4 is the lowest. Hence, M_2 has the highest marginal likelihood. M_3 has the shortest interquartile range and, therefore, variability (a). DIC (b) and WAIC (c). For the log marginal likelihood, higher values are preferred, while for the deviance information criterion (DIC) and widely applicable information criterion (WAIC), smaller values are preferred. All techniques select the data-generating model.

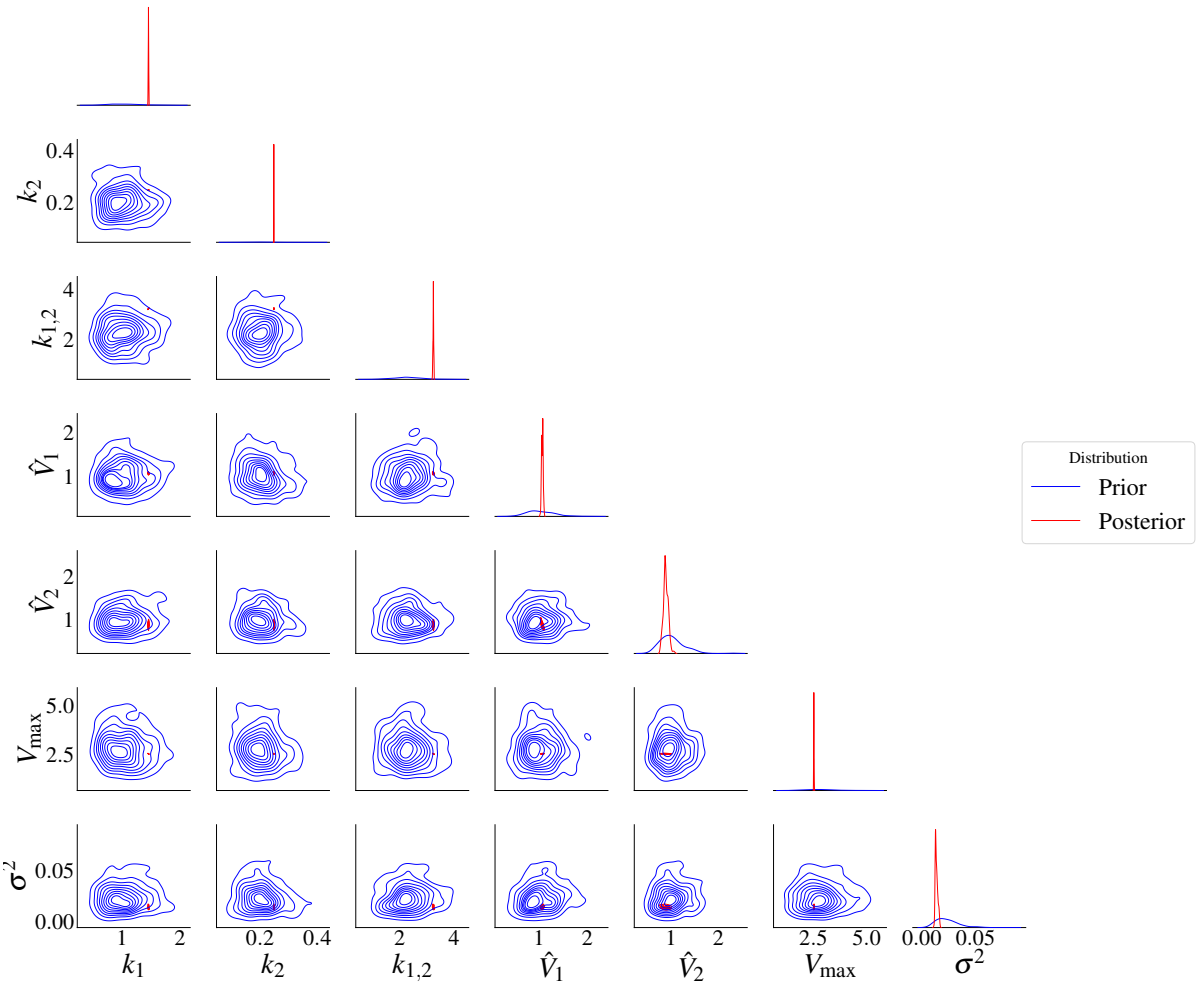


Figure 2.8: Prior and posterior distributions for model M_2. It is difficult to see the correlations due to the high difference in variance between the prior and posterior distributions. The red represents the posterior distributions and the blue the prior distributions. The posterior distributions have contracted compared to the priors.

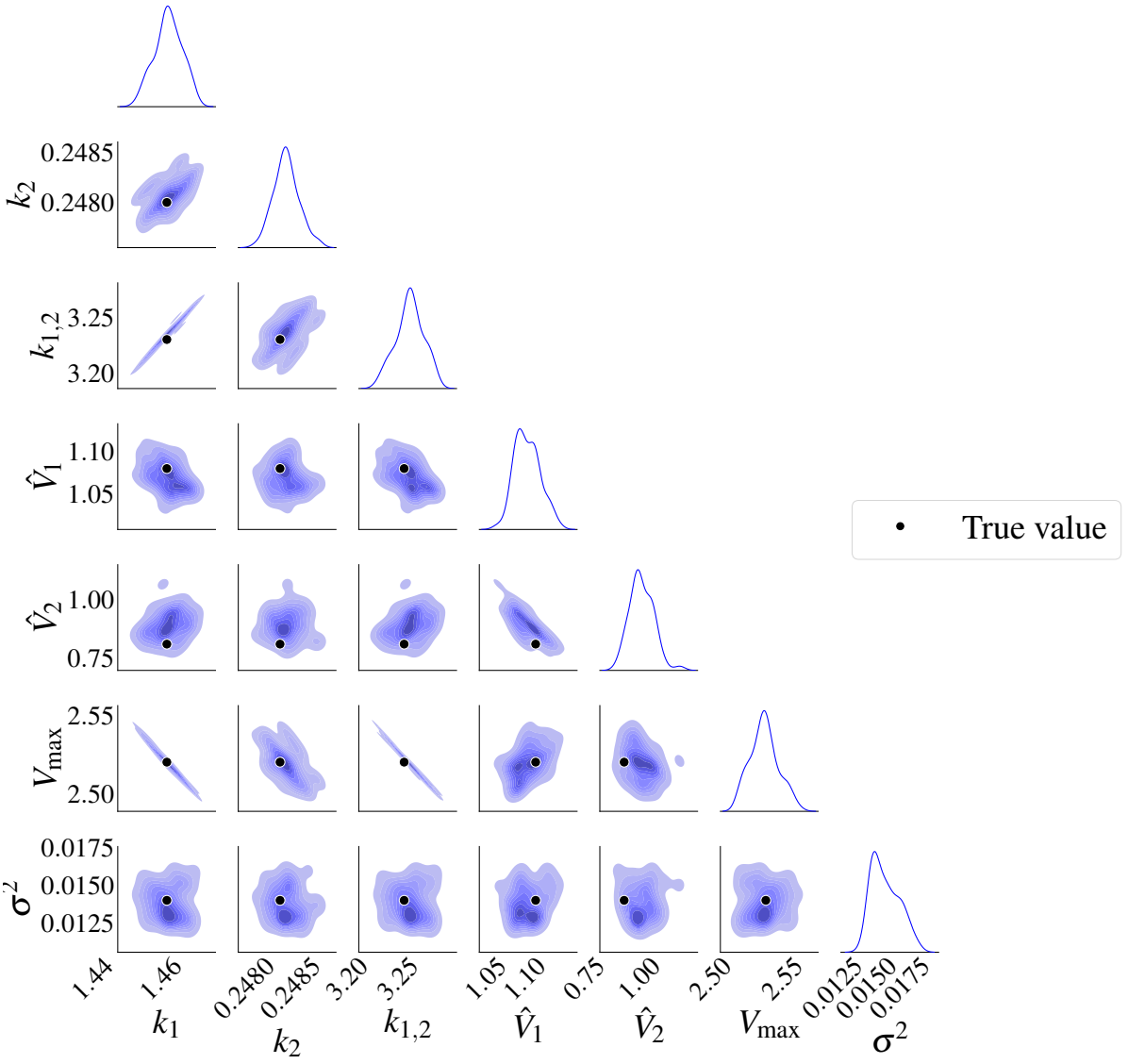


Figure 2.9: Posterior distributions for model M_2. There is a high correlation between k_1 and $V_{\max}$, $k_{1,2}$ and k_2, $k_{1,2}$ and $V_{\max}$. The marginal posterior distributions are on the diagonal. The black dots represent the true parameters used in the data generating process.

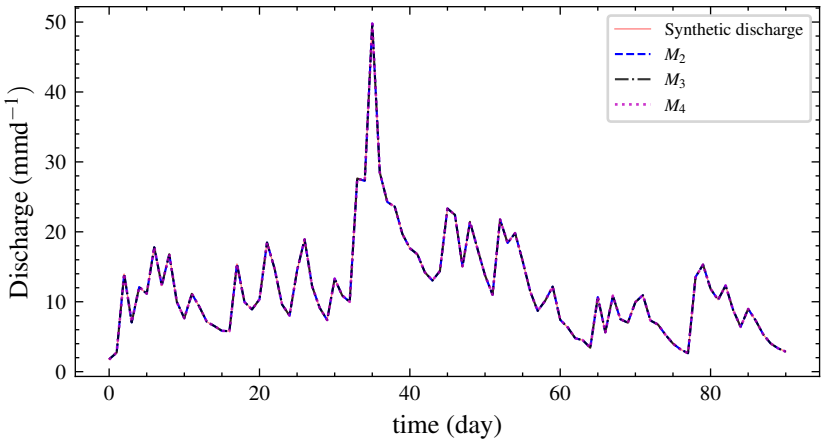


Figure 2.10: Plot of the mean discharge data generated from the posterior predictive distribution of each model for experiment two. It is difficult to choose one model by inspection as they all fit the data well. However, the BF implicitly penalises the unnecessarily complex models M_3 and M_4 and correctly selects M_2.

Posterior predictive plots might not tell us if the chosen model fits the data well, especially for dense data sets. Therefore, formal posterior predictive tests based on the discrepancy measure are needed. Like most statistical tests, the results will depend on the type of discrepancy measure or the test statistics. Carefully choosing such discrepancy measures is crucial. For example, we may test whether the model can predict peak discharge values, which would require a different discrepancy measure than if the aim of our analysis was to predict the mean values. Hence, we suggest using formal posterior predictive tests and graphical posterior predictive checks as in this study.

The PPP is 0.51, which means that the model has good predictive performance. This is expected for synthetic data. Values further from 0.50 indicate a model mismatch with the data. Values closer to zero indicate that the model predictions are lower than the observed data. In contrast, values closer to one point that predictions are higher than observed data. A plot of the autocorrelations of predicted versus synthetic observed data is shown in Fig. 2.11. The proportion of values above the 45^o line is the PPP. We also calculated PCPPP for the selected model and got a value of $0.64 > 0.05$, which implies the model can generate the data. The PCPPP is calibrated based on the prior predictive distribution and is uniformly distributed. Thus, it has the same interpretation as a classical p-value.

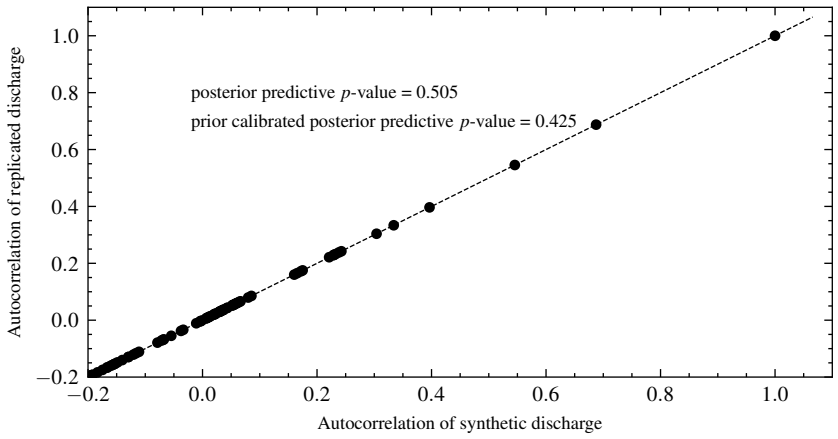


Figure 2.11: Autocorrelation of the replicated versus observed synthetic discharge data. The posterior predictive p-value is the proportion of observations above the 45° line. The autocorrelation of the first point is 1, which isolates it from the other observations.

Experiment two with data generated from the three-buckets model M_3

For the second experiment, the data model is M_3. The model M_3 has three more parameters than M_2 and three fewer parameters than the model M_4. The priors for model M_2 and M_3 are shown in Table 2.3. The data in this experiment was also generated to follow the same trend as the observed data. All models were fitted to the data, and inference is based on 20,000 posterior samples with a burn-in of 5,000. As explained above, convergence was checked using IAT and Geweke diagnostics. The posterior estimates are in Table 2.5. Although the error term is small for all models, M_2 has a higher value than the other two models, suggesting that it may not have the right complexity. Fifteen marginal likelihoods were also calculated for each model in parallel. The mean log marginal likelihood is presented in Table 2.5. The results are also shown in box plots in Fig. 2.12. The box plots reveal that M_3 has the highest median log marginal likelihood, and M_2 the lowest. There is decisive evidence in favour of model M_3, the expected result.

Following the recommendations in (Burnham & Anderson, 2002) for interpreting information theoretic criteria, a difference of 4 to 7 suggests a weak preference for a model and a difference of at least 10 suggests strong preference for a model. Consequently, the DIC and the WAIC do not suggest a strong preference for the true model (M_3) over the richer model M_4. The WAIC shows possible weak evidence in favour of M_3 over M_4, but we note that the error bar in Fig. 2.12 for WAIC M_4 indicates substantial uncertainty in the estimate. In this case then the BF decisively selects the data generating model M_3 where the information theoretic criteria fail to do so. This example alone is clearly not proof that the BF is always superior to WAIC or DIC, but it suggests that there are cases in which BF succeeds and information theoretic criteria can fail. The success of the BF of course comes with a significantly higher computational cost.

As in the experiment one, a hydrograph from the posterior predictive distribution is shown in Fig. 2.13. From the hydrograph, we cannot determine the best model through visual inspection since all the models fit the data equally well. Therefore, we require a formal model selection technique such as the BF.

Table 2.5: True value, posterior mean with 95 % credible intervals of parameters and log marginal likelihood of models for experiment two. M_3 the true model has the highest log marginal likelihood. The DIC and WAIC are also included.

parameter	true value	M_2 (95 % CI)	M_3 (95 % CI)	M_4 (95 % CI)
k_1	1.091	1.109 (1.104, 1.113)	1.090 (1.084, 1.097)	1.089 (1.081, 1.095)
k_2	0.188	0.207 (0.206, 0.207)	0.172 (0.160, 0.190)	0.160 (0.129, 0.174)
k_3	0.208	-	0.208 (0.207, 0.208)	0.241 (0.196, 0.265)
k_4	0.000	-	-	0.208 (0.207, 0.208)
$k_{1,2}$	1.675	1.772 (1.759, 1.786)	1.648(1.613, 1.693)	1.628 (1.552, 1.670)
$k_{2,3}$	1.050	-	1.520 (1.070, 1.781)	1.102 (0.921, 1.400)
k_{34}	0.000	-	-	1.861 (1.105, 2.749)
$\hat{V}_1$	1.317	1.263 (1.224, 1.325)	1.302 (1.242, 1.346)	1.246 (1.181, 1.282)
$\hat{V}_2$	0.936	1.758 (1.622, 1.914)	0.977 (0.733, 1.167)	0.599 (0.474, 0.761)
$\hat{V}_3$	0.910	-	0.856 (0.696, 1.103)	0.731 (0.459, 0.827)
$\hat{V}_4$	0.000	-	-	0.576 (0.433, 0.954)
$V_{\max}$	3.048	2.929 (2.910, 2.948)	3.081 (3.026, 3.127)	3.106 (2.999, 3.149)
σ^2	0.024	0.027 (0.024, 0.030)	0.023 (0.020, 0.027)	0.023 (0.022, 0.027)
$\log p(y\|M)$	-	161.586	173.845	148.060
DIC	-	-401.612	-427.913	-426.127
WAIC	-	-394.247	-420.380	-417.174

-: The parameter is not included in the model.
σ^2: error term.
log$p(y|M)$: log marginal likelihood.

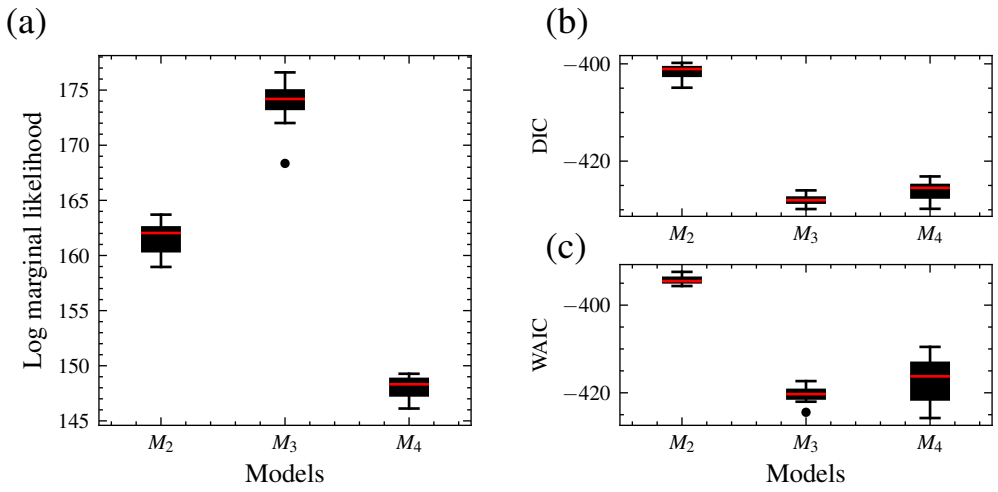


Figure 2.12: Distribution of the log marginal likelihood, DIC and WAIC for 15 different runs each with different initial parameter values. M_3, the data generating model has the highest median log marginal likelihood (a), while M_4 has the lowest. M_4 has the highest number of parameters, while M_2 has the least. DIC (b) and WAIC (c). For the log marginal likelihood, higher values are preferred, while for the DIC and WAIC, smaller values are preferred. The log marginal likelihood selects the data-generating model, while DIC and WAIC do not have any preference for model M_3 and M_4.

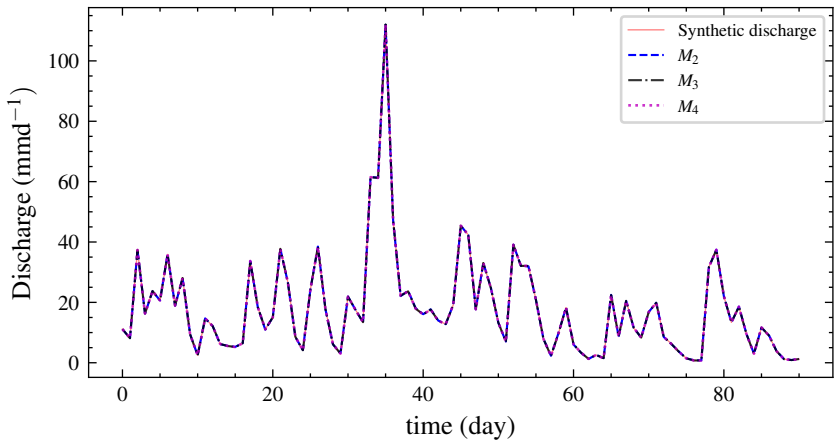


Figure 2.13: Plot of the mean discharge data generated from the posterior predictive distribution of each model for experiment three. It is difficult to choose one model by inspection as they all fit the data equally. The BF implicitly penalises the unnecessarily complex model M_4 and correctly selects M_3.

2.3.3 Real data experiment

This section uses real-world discharge data for Magela Creek in Australia. For each model, 10 chains of the REpHMC were run as in the previous examples. We obtained 4000 posterior samples per chain, discarding the first 1000 as burn-in. The trace plots showed no indication of non-stationarity of the Markov chain, and both Geweke diagnostics and IAT supported convergence. The Z-statistic, p-value, and IAT are shown in Table 2.7. All p-values are greater than 0.05, indicating no significant difference in the means of earlier and later posterior samples and no evidence against convergence. The null hypothesis states that the mean of the earlier and later posterior samples are equal. Furthermore, the IAT is less than $N/50$ for all parameters, indicating well-mixed and stationary chains, where N represents the number of posterior samples. Smaller values of IAT indicate that fewer samples are needed to obtain an independent sample in the Markov chain.

Since we do not use an objective Bayesian approach, we used two sets of priors, where the second set is a sensitivity analysis. The first set of priors has higher variances for some parameters and is less informative than the second set (Table 2.6). It is common practice to try different priors and to check if the parameter estimates change with different priors. This is known as prior-sensitivity analysis. The models converge faster with the second set of priors. The first set of priors (Table 2.3) is the same as in the previous sections. For the second set of priors, we used lognormal priors with lower variances for some parameters compared to the first set of priors. The mean values used for the priors are also different from those of the first set of priors. The prior to the error term remains unchanged.

Table 2.6: Second set of priors. LN is the lognormal distribution and IG is the inverse Gamma distribution

Parameter	Prior distribution
k_1	LN(0.8, 0.25)
k_2, k_3, k_4	LN(0.2, 0.25)
k_{12}, k_{23}, k_{34}	LN(0.6, 0.25)
$\hat{V}_1$, $\hat{V}_2$, $\hat{V}_3$, $\hat{V}_4$, $V_{\max}$	LN(0.0, 0.25)
σ^2	IG(5.0, 0.1)

We checked the precision of our chosen model by comparing predicted and observed discharges using a posterior predictive check based on a second set of priors. The hydrograhs

for all three models are in Fig. 2.16. The plots of the predicted and observed autocorrelations with PPP are in Fig. 2.17. The PPP is 0.444 which is not too close to 0.5 and the PCPPP is 0.639. Hence, one can conclude that the model fits the data based on autocorrelation. Instead of autocorrelation, another metric could be used for the posterior predictive check depending on the objective of the model. The NSE for the chosen model is 0.526 and the KGE is 0.705. This means that the model performs better than using the mean observed discharge. Knoben et al. (2019) found that the KGE is $<$ -0.41 when the model performs poorer than the mean observed discharge. The marginal posterior distributions for the model M_4 are shown in Fig. 2.14. We have also presented the posterior distributions of the parameters in model M_3 in Fig. 2.15. There is no noticeable correlation between parameters when real-world discharge data is used. However, $V_{\max}$ plays a major role in the dynamics of the model. A more realistic prior for $V_{\max}$ based on the soil physics of Magela Creek Australia will reduce the model error.

The results of the second set of priors are in Table 2.8. The selected model did not change when we used diffuse priors. The error in the second set of models is lower than in the first set. The model M_2 is always preferred while M_4 is always the least supported by the data. The error term, its precision, effective sample size (ESS), and the number of parameters influence the marginal likelihood.

We also applied two fully Bayesian information criteria, DIC and WAIC. Unlike the BF, there is no clear model choice for the information criteria. The difference in DIC or WAIC between M_2 and M_3 is less than 1 which means we do not have reason to choose one model over the other.

The RWM, NUTS and MALA were also applied to all the three models with real world data. Even the other gradient-based algorithms NUTS and MALA could not sample the parameter space. Attempts to improve algorithms by trying various values for the initial step size in the case of NUTS and the step size for MALA did not make any difference. This further confirms the fact that combining replica exchange with an algorithm improves mixing and convergence.

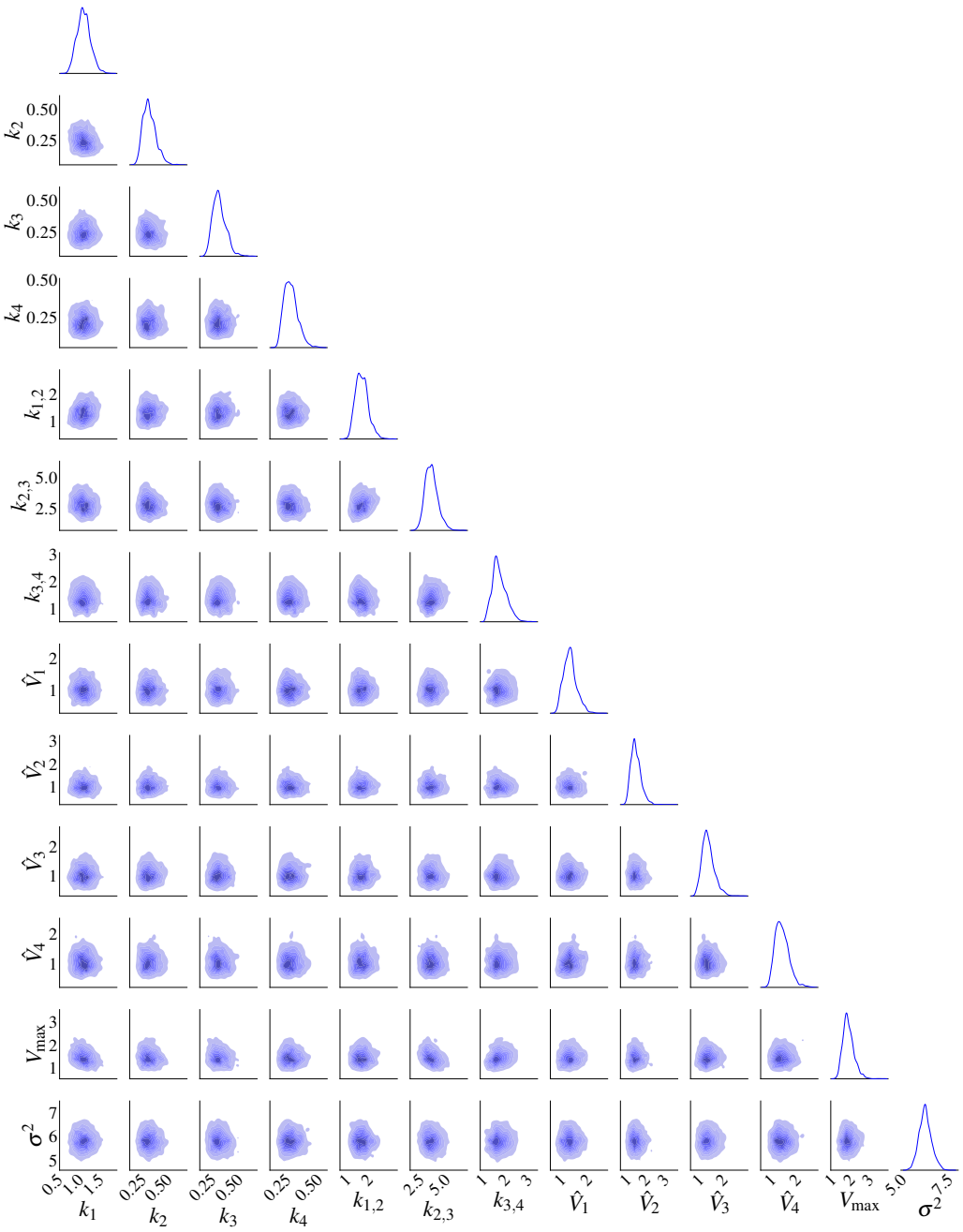


Figure 2.14: Posterior distributions of the 13 parameters for model M_4 using the second set of priors. There is no obvious correlation between the parameters. The marginal posterior distributions are on the diagonal.

Hydrograph of model M_2

Based on the hydrograph Fig. 2.16, most of the model predictions are very close to the observed discharge and within 50 % pointwise credible intervals. However, two peaks are not captured in the model. The first peak discharge period was from 04-02-1980 to 05-02 1980. The observed precipitation during this period is 41.4 mmd^{-1} to 122 mmd^{-1} on 04-02-1980 and 05-02-1980 re-

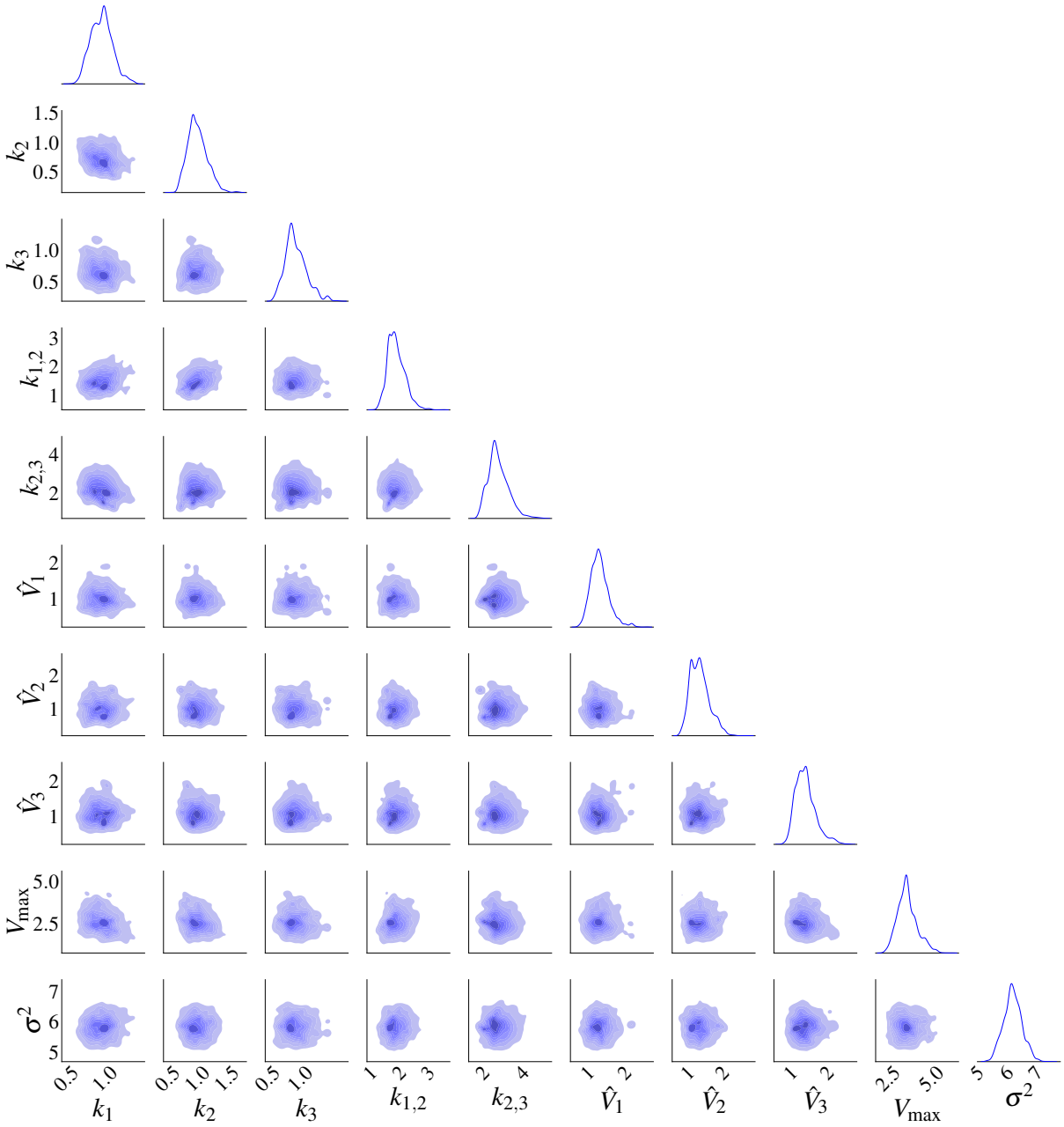


Figure 2.15: Posterior distributions of the 10 parameters of model M_3 based the second set of priors. There is no pronounced correlation between the parameters. The marginal posterior distributions are on the diagonal.

spectively. The observed discharge on these days is 62.09 mmd^{-1} and 21.82 mmd^{-1} respectively. It is illogical that the discharge is reduced with similar weather conditions. The second peak event occurred from 19-03-1980 to 20-03-1980. The precipitation on 19-03-1980 was 39.5 mmd^{-1}, and on 20-03-1980, it was 28.3 mmd^{-1} with potential evapotranspiration similar to other days. This observed discharge is irrational because the discharge from 27-02-1980 (0 mmd^{-1} of precipitation) to 29-02-1980 (22 mmd^{-1} of precipitation) is 18.3 mmd^{-1}. An alternative explanation

Table 2.7: Convergence diagnostics for real-world data. Z-statistic, p-value and IAT. The null hypothesis is that the mean of earlier posterior samples is the same as that of later posterior samples in a Markov chain. All p-values are above 0.05, indicating no significant difference in the mean of earlier and later posterior samples and no evidence against convergence. The IAT is the number of samples required to obtain an independent sample in the Markov chain and smaller values are preferred.

	Model					
	M_2		M_3		M_4	
parameter	Z-statistic (p-value)	IAT	Z-statistic (p-value)	IAT	Z-statistic (p-value)	IAT
k_1	0.029 (0.977)	8.489	-0.169 (0.866)	5.561	-0.001 (0.999)	4.811
k_2	0.631 (0.528)	3.254	0.520 (0.603)	15.302	0.221 (0.825)	16.99
k_3	-	-	-0.432 (0.666)	14.723	0.137 (0.891)	9.892
k_4	-		-	-	-0.371 (0.710)	8.542
$k_{1,2}$	0.136 (0.892)	21.421	0.423 (0.672)	22.547	0.358 (0.720)	12.855
$k_{2,3}$	-	-	0.399 (0.690)	20.578	-0.253 (0.800)	21.233
$k_{3,4}$	-	-	-	-	0.291 (0.771)	9.495
$\hat{V}_1$	-0.801 (0.423)	29.976	0.084 (0.933)	7.650	0.037 (0.970)	11.571
$\hat{V}_2$	-0.809 (0.419)	40.986	-0.015 (0.988)	8.317	0.045 (0.964)	20.099,
$\hat{V}_3$	-	-	-0.226 (0.821)	15.710	0.264 (0.792)	8.548
$\hat{V}_4$	-	-	-	-	-0.402 (0.688)	12.131
$V_{\max}$	-0.146 (0.884)	15.897	-0.184 (0.854)	5.786	0.032 (0.975)	3.953
σ^2	$<$ -0.0001(1.000)	9.092	0.018 (0.985)	1.761	0.001 (1. 000)	2.167

for the mismatch in peak discharge could be that the field capacity of the soil changed during these periods and is not captured in our models.

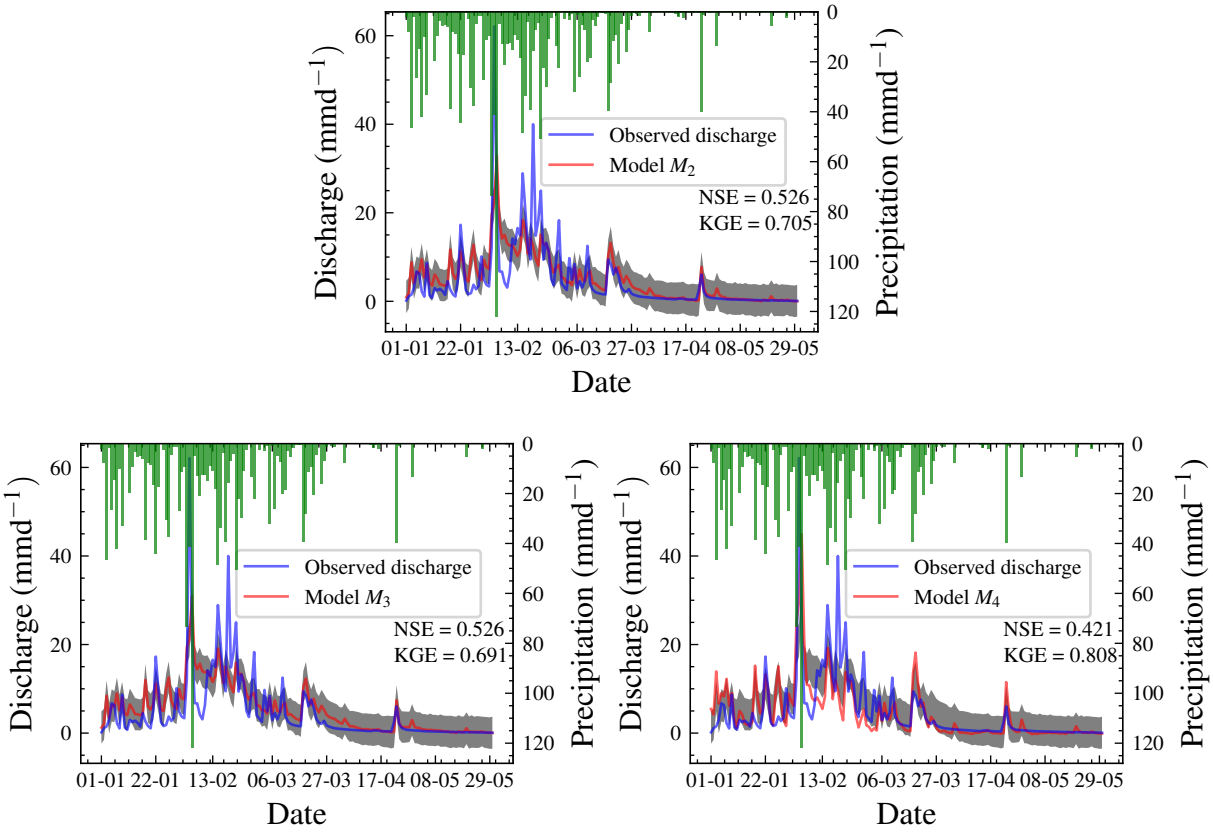


Figure 2.16: Hydrographs for all three models. Models M_2 and M_3 are not visually distinguishable. The results are better than the prior predictive check shown in Fig. 2.6, where most predictions are further from the observed data.

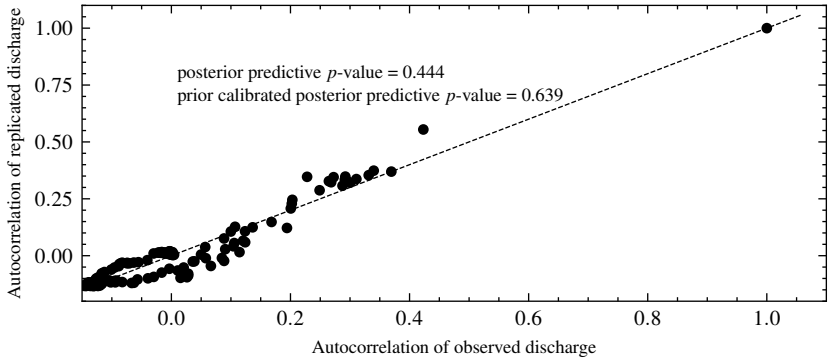


Figure 2.17: Autocorrelation of replicated versus observed data for model M_2. The posterior predictive p-value is the proportion of observations above the 45° line.

Table 2.8: Posterior summary statistics and log marginal likelihood for models with the second set of priors. Model M_2 is the preferred over M_3 based on the log marginal likelihood. The difference in value between model M_2 and M_3 is less than 1 for both the DIC and the WAIC, so there is no preference between the two models according to these criteria. For information-theoretic-based approaches, a difference of 7 is necessary for a strong preference for one model. Model M_4 is the least preferred model based on any approach.

	M_2 (95 % CI)	M_3 (95 % CI)	M_4 (95 % CI)
k_1	0.724 (0.517, 0.940)	0.794 (0.0.574, 1.046)	1.169 (0.774, 1.520)
k_2	0.125 (0.081, 0.174)	0.242 (0.155, 0.344)	1.991 (1.192, 2.801)
k_3	-	0.157 (0.096, 0.221)	1.352 (0.720, 1.964)
k_4	-	-	1.067 (0.598, 1.546)
$k_{1,2}$	1.195 (0.838, 1.637)	1.923 (1.105, 2.889)	2.292 (1.367, 3.417)
$k_{2,3}$	-	0.511 (0.380, 0.648)	0.728 (0.463, 0.983)
$k_{3,4}$	-	-	0.826 (0.497, 1.136)
$\hat{V}_1$	1.030 (0.548, 1.530)	1.029 (0.566, 1.457)	1.140 (0.032, 2.893)
$\hat{V}_2$	1.017 (0.593, 1.549)	0.999 (0.582, 1.477)	0.861 (0.048, 2.239)
$\hat{V}_3$	-	0.997 (0.569, 1.523)	0.940 (0.041, 2.325)
$\hat{V}_4$	-	-	1.082 (0.060, 2.768)
$V_{\max}$	1.139 (0.808, 1.474)	0.912 (0.657, 1.201)	0.796 (0.549, 1.057)
σ^2	5.289 (4.694, 5.830)	5.273 (4.739, 5.828)	5.847 (5.212, 6.499)
$\log p(y\|M)$	-506.259	-529.483	-608.181
DIC	940.352	940.397	969.722
WAIC	946.536	946.512	979.932

-: The parameter is not included in the model.
σ^2: error term.
log$p(y|M)$: log marginal likelihood.

2.3.4 Convergence

Model convergence time

In terms of the theoretical complexity, if N is the number of posterior chains, S the number of samples per chain and L the number of leapfrog steps per sample, then there are on the order of NSL likelihood and likelihood gradient evaluations for the algorithm to complete.

In terms of actual performance, all models converge by 3000 samples, even for real-world data. A single replica set runs on single CPU core within a high-performance computer. The model runtime of Gaussian shell examples ranges from 6 seconds for 2 dimensions to 24 seconds for 30 dimensions. Synthetic examples converge in 2 to 4 hours, depending on the parameter's dimension. On the contrary, with real data, the models converge in 6 to 20 hours, depending on

the parameter space and number of temperatures. Models can converge faster with proper tuning of the number of leapfrog steps. The posterior summary statistics like mean of the parameters does not change much with the number of temperatures. The number of temperatures mainly affects the estimate of the log marginal likelihood. With large datasets, REpHMC can be combined with subsampling without replacement to accelerate convergence. The REpHMC converges in minutes if we are interested only in parameter estimation.

Convergence of marginal likelihood

As proposed by Calderhead & Girolami (2009), most studies use ten temperatures. However, it is important to check for convergence of the log marginal likelihood after convergence of the posteriors. We suggest starting from eight temperatures until the marginal likelihood is stable. That is stop when there is very little variation in the the marginal likelihood. This can be visualised by a graph of the marginal likelihood against the number of temperatures. The number of temperatures at which the log marginal likelihood starts to plateau or flatten is the temperature at which it converges. Also, a horizontal line can be drawn at any point to see where most of the values lie or are close to the line, which helps to check for convergence. As observed with the Gaussian shells example, the marginal likelihood is constant from 10 to 12 temperatures. Thus, running beyond 12 temperatures is recommended. The diagnostic plot of the log marginal likelihood for the real-world example shows that it is constant from 10 to 12 temperatures too Fig. 2.18. For the real-world data, we used 45 temperature schedules for each model.

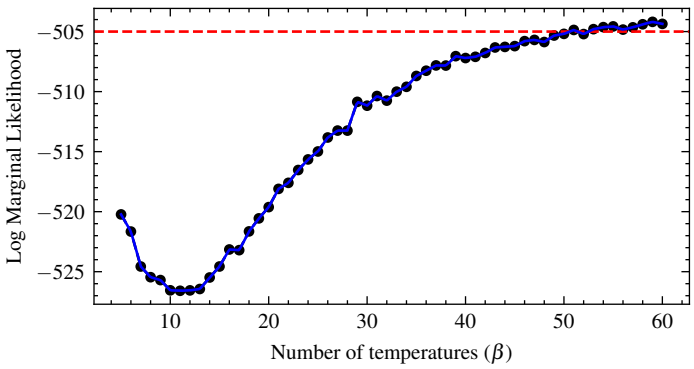


Figure 2.18: Convergence diagnostic of the log marginal likelihood for the two buckets model. The optimal temperature is from 48 when there is very little variation, and the curve begins to flatten. The values almost follow the red line from 45 temperatures.

2.4 Conclusions

We have introduced a methodology for simultaneous Bayesian parameter estimation and model selection. The methodology includes formal model diagnostics, which check for goodness-of-fit and prior data conflict. The method uses a new gradient-based algorithm REpHMC to draw posterior samples, TI for the calculation of marginal likelihood and PCPPP for model diagnostics. The REpHMC and TI were validated on the Gaussian shells example, which is a difficult sampling benchmark problem since it has isolated modes. The REpHMC is effective in sampling the entire parameter space for models with isolated modes. This sets it apart from other gradient-based algorithms such as HMC, NUTS and MALA. Also, we have shown that BF selects the data generating model in two experiments, while DIC and WAIC correctly select the true model in one of two experiments. Also, none of the other mentioned gradient-based algorithms worked when real-world data was used with our developed model. In addition, formal posterior predictive checks have been introduced to determine if a model can accurately predict observed or desired values, such as the minimum or peak discharge. The method was employed to discharge data from Magela Creek in Australia. We also calculated NSE and KGE for the chosen model with real-world data. The framework has been implemented in open-source software TFP which supports most algorithms. The REpHMC can be applied to any hydrological model. Our developed model performed better than using the mean as a predictor for real discharge data. However, the model does not capture peak discharge values. Therefore, some improvements in that regard need to be made.

By combining a gradient-based algorithm HMC and REMC, we get a very powerful algorithm that can sample complex posteriors thanks to the exchange of information between parallel running chains. We have also illustrated that the BF is a reliable Bayesian tool for model selection in contrast to two common Bayesian-based information criteria for model selection.

Future work could combine REMC with NUTS algorithm to automatically tune all parameters in the HMC. Also, introducing subsampling in the case of big data or models with millions of parameters will reduce the inference time. Future work could focus on improving the model goodness-of-fit, as the KGE indicates. Furthermore, one could develop a discrepancy measure for the posterior predictive check to test whether the selected model can capture peak discharge values. On the practical side, this study could be extended to the multi-catchment setting. Also, different types of conceptual hydrological models could be compared using this approach.

Code and data availability

The source code, data, and instructions are available on Zenodo (N. D. Mingo & Hale, 2023) and GitHub at `https://github.com/DamingoNdiwa/hydrological-model-selection-bayes`.

Author contributions.

DNM: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualisation, Writing - original draft, Writing - review & editing. RN: Conceptualization, Formal analysis, Methodology, Writing - review & editing. CL: Conceptualization, Funding acquisition, Formal analysis, Methodology, Supervision, Writing - review & editing. JSH: Conceptualization, Formal analysis, Funding acquisition, Methodology, Project administration, Software, Supervision, Validation, Writing - original draft, Writing - review & editing.

Competing interests

The authors declare no competing interests.

Acknowledgements

We would like to thank Stanislaus Schymanski for his valuable insights which inspired us to undertake this study, for his critical feedback on a draft of this manuscript. This work was funded under the Luxembourg National Research Fund under the PRIDE programme (PRIDE17/12252781) and ATTRACT programme (A16/SR/11254288). The experiments presented in this paper were carried out using the HPC (Varrette et al., 2022) facilities of the University of Luxembourg – see `https://hpc.uni.lu`.

Chapter 3

Bayesian prior impact assessment for dynamical systems described by ordinary differential equations

The content of this chapter is based on the following pre-print.

Mingo, D. N., Hale, J. S., Ley, C. (2024). Bayesian prior impact assessment for dynamical systems described by ordinary differential equations. University of Luxembourg Open Repository and Bibliography [Submitted preprint], `https://hdl.handle.net/10993/61471`

Author contributions.

DNM: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing - review & editing. JSH: Conceptualization, Formal analysis, Funding acquisition, Resources, Methodology, Project administration, Software, Supervision, Validation, Writing - original draft, Writing - review & editing. CL: Conceptualization, Funding acquisition, Formal analysis, Methodology, Supervision, Writing - review & editing.

Abstract

In this study we extend the use of the Wasserstein Impact Measure (WIM) to the problem of assessing prior impact in Bayesian models governed by systems of Ordinary differential equations (ODEs) with moderate (5 to 10) parametric dimension. First, we utilise algorithms from computational optimal transport to compute the WIM in moderate parametric dimensions. Second, we propose a new prior scaled Wasserstein Impact Measure (sWIM) measure which gives a relative sense of distance, easing with interpretation of the WIM for understanding the impact of the prior on the resulting inference. We show numerical computation and interpretation of the WIM and sWIM for a Lotka-Volterra predator-prey model calibrated against the Hudson Bay Company dataset and a compartment epidemiological model calibrated against first-wave COVID-19 data from Luxembourg.

3.1 Introduction

A strength of Bayesian statistics is the ability to easily incorporate prior information, such as historical or expert knowledge, into a parameter inference problem. However, the inherent price to pay for this flexibility is that the choice of prior may have a strong impact on the subsequent parameter inference. Consequently a crucial component of Bayesian analysis is the choice and justification of a prior, along with understanding the impact of prior choice on the resulting parameter estimates and model output.

Bayesian methods are widely used for solving parameter identification problems of continuous time dynamical systems modelled by Ordinary differential equations (ODEs) (Girolami, 2008). Fields of study where ODEs are a prominent methodology for constructing models include epidemiology (Gibson et al., 2023; Kemp et al., 2021), hydrology (Machac et al., 2016) and mechanics (Awrejcewicz, 2014). These models typically contain 5 to 10 unknown parameters that have a critical effect on the overall behaviour of the system, and so it is necessary to identify these parameters using observed data before the model can be put to use.

We now discuss some of the issues particular to solving parameter identification problems involving ODEs. Firstly, the observed data to calibrate the model is often sparse in the sense that there are only observations available at a limited number of points in time, and for each time point, usually only a single (noisy) observation. Additionally there are usually only observations on a limited subset of the system states. In some cases, we may only be able to observe a proximal

function of the system states, for example a weighted sum. Due to this sparsity it is often the case that the data can only weakly constrain the parameters, leading to inference problems that are inherently ill-posed. Aside from this notion of data sparsity, the parameters themselves are usually constrained *a priori* by fundamental physical considerations, e.g. positivity, or by a large body of expert knowledge from previous related studies on perceived 'sensible' values.

Because of these issues, Bayesian methods, precisely because of the inclusion of a prior, are a natural choice for parameter identification problems involving ODEs. Specifically, the inclusion of prior knowledge can turn a problem that is inherently ill-conditioned due to data sparsity, into a well-conditioned problem with a reasonable solution. Furthermore, the prior gives a way to directly encode information about physical constraints on the parameters and/or to incorporate historical data. Because of this priors used with ODEs are rarely of the non-informative or objective type (Berger et al., 2015; Ghosh, 2011), with a preference for weakly informative (Gelman et al., 2017; Gelman, 2006) or strongly informative priors (Gibson et al., 2023; Lai et al., 2021; Calderhead & Girolami, 2011). We also remark that in these problems we are almost always working far from the regime where Bernstein–von–Mises type results may be applicable. Furthermore it has been shown in e.g. (Golchi, 2019; Gelman et al., 2017; Gelman, 2006) that weakly and strongly informative priors can lead to reduced computational cost over non-informative priors when exploring the posterior distribution using MCMC. Due to this widespread use of informative priors, it is desirable to develop quantitative methods to assess how the choice of prior impacts posterior parameter estimates. For example, do the priors have similar impact or does one have a higher impact on the posterior?

Before proceeding to our contribution, we take a moment to review the existing literature on assessing prior impact in a general Bayesian model context - we are not aware of any studies that discuss quantitative prior impact methods in the context of ODE systems. A common approach is to recalibrate the model with slightly different priors and judge how it impacts the resulting inference (Stefan et al., 2022; Pedroza et al., 2018; Schmidli et al., 2014; Nur et al., 2009). Such an approach is qualitative as it tells us if different priors lead to different posterior statistics, but there is typically no quantitative measure of the difference between the resulting posteriors.

Quantitative measures of posterior difference can be obtained by calculating the discrepancy between distributions as measures of informativeness. Popular discrepancy measures include the KL divergence, mean square error, Wasserstein distance and Hellinger distance. The KL divergence has been used to calculate the prior effective sample size (ESS) (Morita et al., 2008) as a measure of prior impact. The ESS is the number of observations with the same amount

of information as the prior. This measure can suffer from being over-estimated for mixture and multimodal distributions. The mean square error has been used to determine the effective current sample size (Wiesenfarth & Calderazzo, 2020), which is the number of observations that have to be added or subtracted from the prior to obtain the same inference as a baseline prior. The effective current sample size might change if the mode or median is used instead of the mean. Jones et al. (2022) used the Wasserstein distance to calculate the mean observed effective sample size (MOPESS).

Discrepancy measures have also been used for prior assessment outside of determining the effective sample. Tang et al. (2016) used the KL divergence to quantify prior informativeness in hydrology. Ghaderinezhad et al. (2022) introduced the Wasserstein Impact Measure (WIM) which captures the Wasserstein distance between two posterior distributions resulting from two different priors. They calculated the WIM for univariate and bivariate parameters. Weiss (1996) used the chi-square divergence as an interpretable measure of prior sensitivity. Roos et al. (2015) applied the Hellinger distance to quantify the prior impact for hierarchical models. Gustafson (1996) employed the L^2 norm to evaluate the sensitivity of prior perturbations to posterior expectations. The above divergences and distances (Ali & Silvey, 1966) have trade offs and are generally chosen based on a combination of convenience (Roos et al., 2015), their suitability for the problem at hand and the availability of a robust algorithm and software implementation for their calculation. The KL divergence and chi-square divergence are not symmetric, and are consequently not metrics/distances. The KL divergence is undefined if the intersection of the support of the distributions is the empty set. The Hellinger distance is bounded between zero and one and is a probability metric. The Wasserstein distance is symmetric, computable between discrete and continuous probability distributions, takes into account the geometry of the parameter space (Panaretos & Zemel, 2019) and is a properly defined metric. For these reasons, we focus now on the Wasserstein metric.

3.1.1 Contributions

We make two main contributions in this paper. Firstly, we extend the WIM proposed in Ghaderinezhad et al. (2022) to Bayesian models involving ODEs. This extension is enabled by recent advances in computational optimal transport (Cuturi, 2013; Cuturi et al., 2022; Cuturi & Doucet, 2014) allowing for the efficient estimation of the Wasserstein distance in moderate dimensions.

Secondly, we propose a new prior impact measure, which we call the prior scaled Wasserstein

Impact Measure (sWIM), that improves on the WIM by endowing it with a relative sense of scale. To fix ideas, we describe the sWIM and WIM now in words. Given a baseline prior and a prior of interest, and the two induced posteriors, the sWIM scales the Wasserstein distance between the posteriors (the definition of the original WIM) by the distance between the priors. This scaling is inspired by the one developed in (Roos & Held, 2011; Roos et al., 2015) which gives a relative change of scale to the Hellinger distance when used as a prior impact measure for priors with perturbed parameters. We show that this sense of scale helps overcome some of the difficulties interpreting the WIM, particularly whether a WIM for a given problem is 'big' (impactful) or 'small' (not impactful). We also introduce the marginal sWIM as the sWIM for each parameter with a different prior from the baseline instead of for all parameters.

3.1.2 Outline

An outline of this paper is as follows. In Section 3.2 we give an overview the core components of the methodology for calculating the sWIM for ODE systems. Then in Section 3.3 demonstrate and discuss the methodology on a Lotka-Volterra predator-prey ODE model and a Susceptible-Exposed-Infected-Removed (SEIR) epidemiological model, before concluding in Section 3.4.

3.2 Methodology

In this section, we begin by introducing the basic notion of a Bayesian inference problem involving ODEs. We then discuss the Wasserstein distance from an optimal transport perspective (Villani, 2009) and then introduce the Sinkhorn algorithm that we use to calculate it. We then discuss the WIM and introduce the new sWIM prior impact measure.

3.2.1 Bayesian inference for ordinary differential equations

To set ideas we briefly review Bayesian inference in the context of a model involving the solution of an ODE.

Bayesian inference is based on Bayes' theorem which states that the posterior distribution is proportional to the product of the data likelihood and the prior distribution. In the subjective Bayesian framework, the prior reflects the practitioner's beliefs, expert knowledge or any other historical information one may have about the parameters of interest. The likelihood indicates how likely it is for a parameter value to have generated the data. For a given dataset $y \in \mathbb{R}^n$

and parameter vector $\theta \in \mathbb{R}^p$, Bayes' theorem can be stated as

$$p(\theta|y) = \frac{f(y|\theta)p(\theta)}{p(y)}, \tag{3.1}$$

where $p(\theta|y)$ is the posterior distribution, $p(\theta)$ the prior distribution, $f(y|\theta)$ the likelihood function and $p(y)$ a normalising constant.

An ODE describes how the state $z(t)$ changes over time $t \in (0, T]$ and can be written in standard form as

$$\frac{\mathrm{d}z}{\mathrm{d}t} = F(t, z, \theta), \tag{3.2a}$$

$$\hat{z} = z(t = 0), \tag{3.2b}$$

where $\hat{z} \in \mathbb{R}$ is the initial condition and $F : (0, T] \times \mathbb{R} \times \mathbb{R}^p \to \mathbb{R}$ a known function particular to the modelled system. As there is usually no closed-form solution of Eq. (3.2) we resort to numerical methods to find a solution, see e.g. (Kidger et al., 2020) for details.

To link Eq. (3.2) with Eq. (3.1), without loss of generality, we assume the following data generating model for y

$$y_i \mid \theta \sim N([G_z(\theta)]_i, \sigma^2), \quad i = 1, \ldots, n, \tag{3.3}$$

where $G_z(\theta) : \mathbb{R}^p \to \mathbb{R}^n$ requires the solution of Eq. (3.2) for z at a fixed θ and the subsequent evaluation of z at a set of n observation points in time $(t_1, t_2, \ldots, t_n)$ with each $t_i \in (0, T]$. With the additional specification of priors θ, the Bayesian posterior in Eq. (3.1) is fully defined and can be explored using standard techniques e.g. MCMC (see e.g. Gelman et al., 2014; Hoffman & Gelman, 2014).

3.2.2 Wasserstein distance

The definition of the p-Wasserstein distance (W_p) between two probability measures μ, ν defined on the space $\mathcal{X}$ is

$$W_p(\mu, \nu) = \inf_{\pi \in U(\mu,\nu)} \left(\int_{\mathcal{X} \times \mathcal{X}} ||x - y||^p d\pi(x, y) \right)^{1/p}, \quad p \geq 1, \tag{3.4}$$

where $U(\mu, \nu)$ is the set of joint probability measures on $\mathcal{X} \times \mathcal{X}$ (Villani, 2009). Calculating the Wasserstein distance becomes non-trivial in moderate to high dimensions due to the ill-posed nature of the squared Euclidean distance (Cuturi et al., 2023). The Wasserstein distance is

the minimum amount of work required to reconfigure the mass of one distribution into another (Panaretos & Zemel, 2019).

3.2.3 Wasserstein impact measure

Let P_0 and P_1 be posteriors induced from a baseline prior p_0 and a prior of interest p_1, respectively, with all other factors defining the Bayesian problem (likelihood, data etc.) kept the same. The WIM from (Ghaderinezhad et al., 2022) is defined as

$$\mathrm{WIM}\left(p_0, p_1\right) = W_2\left(P_0, P_1\right). \tag{3.5}$$

In order to overcome the difficulty in interpreting the WIM, we take inspiration from (Roos et al., 2015) that scaled the Hellinger distance between posteriors by the distance between priors with applications to hierarchical Bayesian models. We thus propose to divide the Wasserstein distance (the original WIM) between the two posteriors by the Wasserstein distance between the two priors, resulting in

$$\mathrm{sWIM}\left(p_0, p_1\right) = \frac{\mathrm{WIM}\left(p_0, p_1\right)}{W_2\left(p_0, p_1\right)}. \tag{3.6}$$

The interpretation of our new measure of prior impact is similar to Roos et al. (2015): when $\mathrm{sWIM} < 1$, the distance between posteriors is smaller compared to the distance between priors, while when $\mathrm{sWIM} > 1$, the distance between posteriors is greater than the distance between priors. The interpretation of the sWIM is in Table 3.1 In addition, when $\mathrm{sWIM} \simeq 1$, the differences in prior and posterior are similar and hence not much affected by the data (via the likelihood). The sWIM can also be interpreted as the relative change in the posteriors due to a change in priors. This interpretation is inspired by Roos & Held (2011) for their chi-squared divergence for sensitivity analysis

$$\chi^2(P_0) = \frac{\mathbb{E}[P_0] - \mathbb{E}[P_1]}{\mathbb{E}[P_0]} \tag{3.7}$$

The posterior expectation of the baseline prior is $\mathbb{E}[P_0]$ while the posterior expectation of the prior in question is $\mathbb{E}[P_1]$. The $\chi^2(P_0)$ is interpreted as the relative change in the posterior expectation due to perturbations in the baseline prior.

In summary, our approach to prior impact assessment involves:

1. Choose a baseline prior denoted as p_0 and a prior of interest denoted as p_1.

Table 3.1: Interpretation of the prior scaled WIM (sWIM)

sWIM	weight of impact compared to a baseline
< 1	Priors have similar impact on the posterior.
> 1	Priors have different impact on the posterior.
$\simeq 1$	The differences in prior and posterior are similar and hence not much affected by the data

2. Obtain m samples from the baseline prior and m samples from the prior of interest.

3. Perform Bayesian inference on the two models to obtain m baseline posterior samples from P_0 and m posterior samples from the prior of interest P_1.

4. Compute the Wasserstein distances between the baseline prior samples and those of interest, as well as between the baseline posterior samples and the samples of interest.

5. Finally calculate the sWIM, marginal sWIM using Eq. (3.6) and interpret the resulting quantities.

3.2.4 Discrete optimal transport and Sinkhorn algorithm

For the practical calculation of the sWIM, we make use of techniques from discrete optimal transport. For a full introduction to we refer the reader to (Cuturi & Doucet, 2014; Scetbon et al., 2021), and we use similar notations to these two papers.

Consider two probability measures μ and ν approximated by m weighted discrete samples $X = \{x_1, x_2 \ldots, x_m\}$ and $Y = \{y_1, y_2 \ldots, y_m\}$, respectively

$$\mu = \sum_{i=1}^{m} a_i \delta(x_i), \qquad \nu = \sum_{j=1}^{m} b_j \delta(y_j), \tag{3.8}$$

where δ is the usual Dirac delta function, a is a vector of weights with elements $a_i > 0$ and b is a vector of weights with elements $b_j > 0$.

Let C_{ij} be a cost matrix, which is the pairwise metric between the elements of the samples X and Y. When the metric is chosen as the squared Euclidean distance, as in our case, the cost matrix is

$$C_{ij} = \|x_i - y_j\|^2, \quad i = 1, \ldots, m, \quad j = 1, \ldots, m. \tag{3.9}$$

For two matrices of the same size, A and B, the Frobenius inner product is

$$\langle A, B \rangle = \mathrm{Tr}(A^T B).$$

Then, the squared 2-Wasserstein distance between two discrete probability measures μ and ν is

$$W_2^2(\mu, \nu) = \min_{T \in \mathrm{U}(a,b)} \langle C, T \rangle \tag{3.10}$$

Eq. Eq. (3.10) is the primal formulation of the Wasserstein distance, where T is the joint probability and

$$\mathrm{U}(a, b) \stackrel{\text{def}}{=} \{T \in \mathbb{R}_+^{m \times m} : T^T \mathbb{1}_m = a \text{ and } T\mathbb{1}_m = b\}.$$

contains all possible joint probabilities.

Cuturi (2013) introduced entropic regularized optimal transport. This approach regularizes the optimal Wasserstein distance with an entropy term. The entropy $H(T)$ is defined as

$$H(T) = -\sum_{ij} T_{ij}(\log T_{ij} - 1).$$

The entropy regularized 2-Wasserstein distance is defined as

$$W_{2,\varepsilon}^2(\mu, \nu) = \min_{T \in \mathrm{U}(a,b)} \langle C, T \rangle - \varepsilon H(T), \tag{3.11}$$

where ε controls the strength of the penalty, when ε is zero, we recover the original problem. The entropy-regularised optimal transport problem is convex and has a unique solution (Peyré & Cuturi, 2019). Entropy based penalty is inspired by the Schrödinger problem (see Ghosal et al., 2022; Peyré & Cuturi, 2019). The entropy regularized primal and dual formulations are linear programming problems and can be solved with a simple iterative scaling algorithm known as the Sinkhorn algorithm (Cuturi & Doucet, 2014; Scetbon et al., 2021).

3.3 Examples

In this section, we illustrate the proposed prior impact assessment methodology on the Lotka-Volterra predator prey and the SEIR models calibrated against real-world data. By using different sets of priors, we aim to gain insights into how priors impact inference. Specifically, we

seek to answer questions such as whether a prior is more informative than some baseline and, if so, whether its impact on posterior inference is small or large. To do this, we calculate the WIM and sWIM for different sets of priors as well as comparing the parameter estimates and graphical posterior predictive checks. The computational models are constructed using TensorFlow Probability (Dillon et al., 2017) on JAX (Frostig et al., 2018), the NUTS algorithm (Hoffman & Gelman, 2014) is used to obtain posterior samples and Optimal Transport Tools (OTT) (Cuturi et al., 2022) is used to calculate Wasserstein distances with the Sinkhorn algorithm. Complete code to reproduce the results is available in the supplementary material.

3.3.1 Lotka–Volterra model

The Lotka–Volterra model describes predator-prey dynamics in an environment

$$\frac{\mathrm{d}u}{\mathrm{d}t} = \alpha u - \beta uv, \quad \hat{u} = u(t=0), \tag{3.12a}$$

$$\frac{\mathrm{d}v}{\mathrm{d}t} = \delta uv - \gamma v, \quad \hat{v} = v(t=0), \tag{3.12b}$$

where $u(t) > 0$ represents the number of preys at time t, $v(t) > 0$ the number of predators at time t, $\alpha > 0$ is the prey birth rate, $\beta > 0$ links prey mortality to the number of preys and predators, $\delta > 0$ links the increase in predators to the number of predators and preys, and $\gamma > 0$ stands for the predator death rate. The initial prey state is $\hat{u}$ and the initial predator state is $\hat{v}$.

Let the parameter vector $\theta = (\alpha, \beta, \delta, \gamma, \hat{u}, \hat{v})^T \in \mathbb{R}^6$. The data generating model can be constructed as

$$\begin{aligned} z_i^1 \mid \theta &\sim \mathrm{LN}\left([G_u(\theta)]_i, \sigma_u^2\right), \quad \forall i = 1, \ldots, n, \\ z_i^2 \mid \theta &\sim \mathrm{LN}\left([G_v(\theta)]_i, \sigma_v^2\right), \quad \forall i = 1, \ldots, n, \\ y &= \left(z^1, z^2\right)^T \in \mathbb{R}^{2n}. \end{aligned}$$

where the operators G_u and G_v involve the solution of Eq. (3.12a) and Eq. (3.12b), respectively, at n time points distributed between 1845 and 1935, and LN is the log-normal distribution.

We consider hare-lynx data based on historical pelt records of the Hudson Bay Company (Zhibin et al., 2007) which is available in numpyro (Bingham et al., 2019). The pelt records are used as a proxy for the populations of hare and lynx in the environment. Part of this data from 1900 to 1920 has been analysed in (Carpenter, 2018) using a Bayesian approach, and the entire data has been used for model calibration using a frequentist approach in (Xinyu et al., 2015).

Since all involved parameters must be positive, the prior distributions on the parameters are defined as

$$\begin{aligned}\alpha, \gamma &\overset{\text{iid}}{\sim} \text{TN}(1.0, 0.5),\\ \beta, \delta &\overset{\text{iid}}{\sim} \text{TN}(0.05, 0.05),\\ \hat{u}, \hat{v} &\overset{\text{iid}}{\sim} \text{LN}(\ln(10), 1.0),\\ \sigma_u^2, \sigma_v^2 &\overset{\text{iid}}{\sim} \text{LN}(-1, 1),\end{aligned}$$

where TN is the truncated normal distribution. The priors are the same as those used in (Carpenter, 2018). In order to obtain a variety of priors for our prior impact assessment, we perturb the prior distribution for the initial states $(\hat{u}, \hat{v})$ as well as for the error variances (σ_u^2, σ_v^2) to create four distinct sets of priors (p_0, p_1, p_2, p_3) as shown in Table 3.2.

Table 3.2: Priors used for the Lotka–Volterra model. Only the priors on the initial states $(\hat{u}, \hat{v})$ and on the error variances (σ_u^2, σ_v^2) are perturbed. The truncated normal (TN) and the log-normal (LN) distributions are used.

p_0 (Carpenter, 2018)	p_1	p_2	p_3
$\alpha, \gamma \overset{\text{iid}}{\sim} \text{TN}(1.0, 0.5)$	$\alpha, \gamma \overset{\text{iid}}{\sim} \text{TN}(1.0, 0.5)$	$\alpha, \gamma \overset{\text{iid}}{\sim} \text{TN}(1.0, 0.5)$	$\alpha, \gamma \overset{\text{iid}}{\sim} \text{TN}(1.0, 0.5)$
$\beta, \delta \overset{\text{iid}}{\sim} \text{TN}(0.05, 0.05)$	$\beta, \delta \overset{\text{iid}}{\sim} \text{TN}(0.05, 0.05)$	$\beta, \delta \overset{\text{iid}}{\sim} \text{TN}(0.05, 0.05)$	$\beta, \delta \overset{\text{iid}}{\sim} \text{TN}(0.05, 0.05)$
$\hat{u}, \hat{v} \overset{\text{iid}}{\sim} \text{LN}(\ln(10), 1.0)$	$\hat{u}, \hat{v} \overset{\text{iid}}{\sim} \text{LN}(\ln(2), 1.0)$	$\hat{u} \overset{\text{iid}}{\sim} \text{LN}(\ln(15), 1.0)$	$\hat{u} \overset{\text{iid}}{\sim} \text{LN}(\ln(15), 1.0)$
$\sigma_u^2, \sigma_v^2 \overset{\text{iid}}{\sim} \text{LN}(-1, 1)$	$\sigma_u^2, \sigma_v^2 \overset{\text{iid}}{\sim} \text{LN}(-1, 1)$	$\hat{v} \overset{\text{iid}}{\sim} \text{LN}(\ln(6), 1.0)$	$\hat{v} \overset{\text{iid}}{\sim} \text{LN}(\ln(6), 1.0)$
		$\sigma_u^2, \sigma_v^2 \overset{\text{iid}}{\sim} \text{LN}(2.0, 0.2)$	$\sigma_u^2, \sigma_v^2 \overset{\text{iid}}{\sim} \text{LN}(1.0, 0.1)$

We fitted four different models corresponding to the four sets of priors. We obtained the posterior samples using the preconditioned NUTS. A thousand samples were discarded as warm-up and three thousand used for inference. The marginal posterior distributions and pairwise correlations are displayed in Fig. 3.1. We can see high pairwise correlations between some parameters, e.g. in the couples $(\alpha, \hat{u}), (\alpha, \delta)$, and (α, γ). The choice of gradient-based algorithms such as NUTS is crucial to sample the entire posterior distribution, and we show the posterior mean estimates in Table 3.3. The prey birth rate α is more than one unit higher for the baseline than the other prior models. Also, the posterior mean estimates of the initial number of preys $\hat{\mu}$ and initial predators $\hat{\nu}$ are more than two units greater for each model compared to the baseline priors.

Now we calculate the WIM and the sWIM between various pairs of distributions. The results of the WIM are shown in Table 3.4 along with other Wasserstein-2 distances, while those

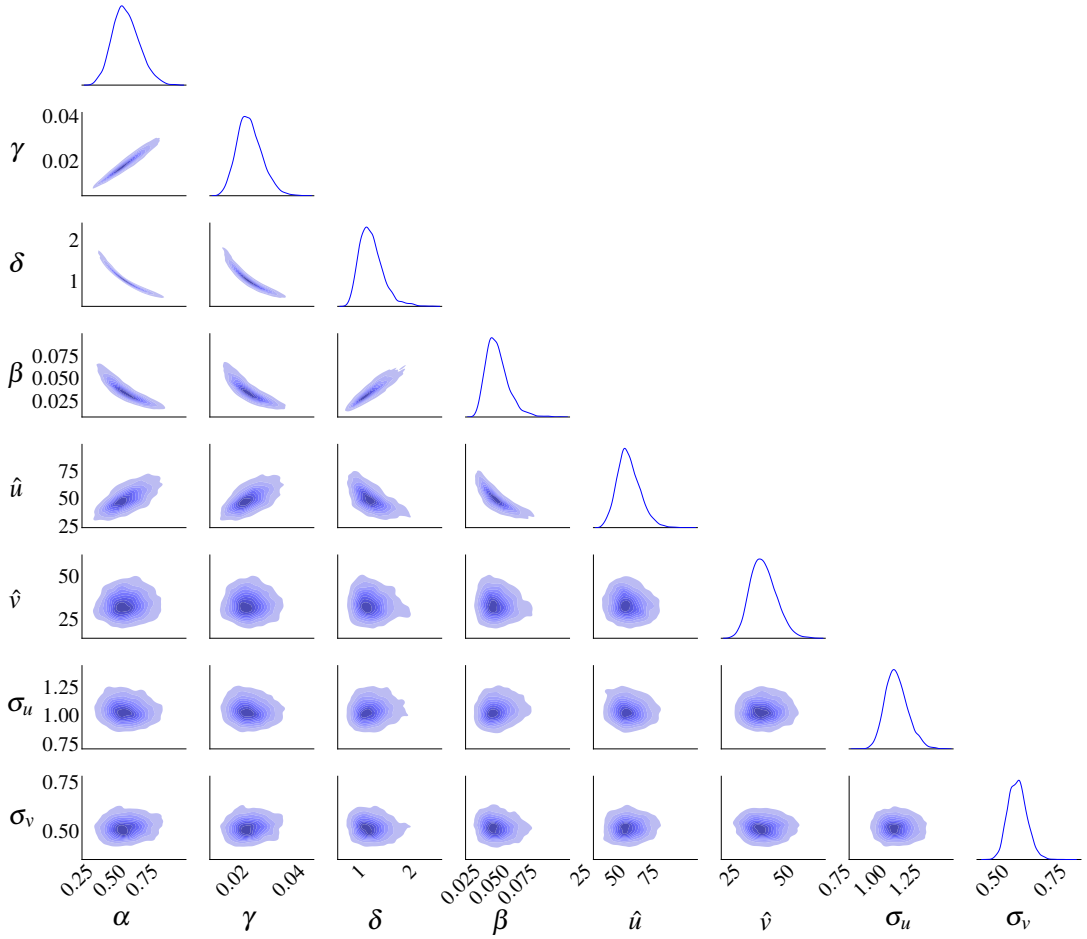


Figure 3.1: Plot of posterior distributions associated to p_0 in the Lotka-Volterra model, with posterior marginal distributions on the diagonal and bivariate distributions for outside the diagonal. One can observe high correlations between some pairs of parameters.

related to the sWIM appear in Table 3.5. We first note that based on both WIM and sWIM the prior p_1 is closest to the baseline. This is in line with the posterior predictive check in Fig. 3.2(a). The sWIM for p_2 and p_3 are greater than one. This means that p_2 and p_3 have different impact on the posterior, in the sense that in each case the posteriors are further from the baseline posterior compared to the distances between priors. The results are consistent with the graphical posterior predictive check, which shows that the predictions are noticeably different from the baseline Fig. 3.2(a). In order to get more detailed information, we compute the marginal sWIM for the initial conditions and error variances and the results are given in Table 3.6. The marginal sWIM is greater than one for the initial condition ($\hat{\mu}$) with priors p_1, p_2 and p_3. Thus, the priors have different impacts on the posterior number of hares compared to the baseline prior. This information is not immediately apparent in the posterior predictive plots Fig. 3.2(a), as p_2 and the baseline prior (p_0) appear indistinguishable. However, it becomes

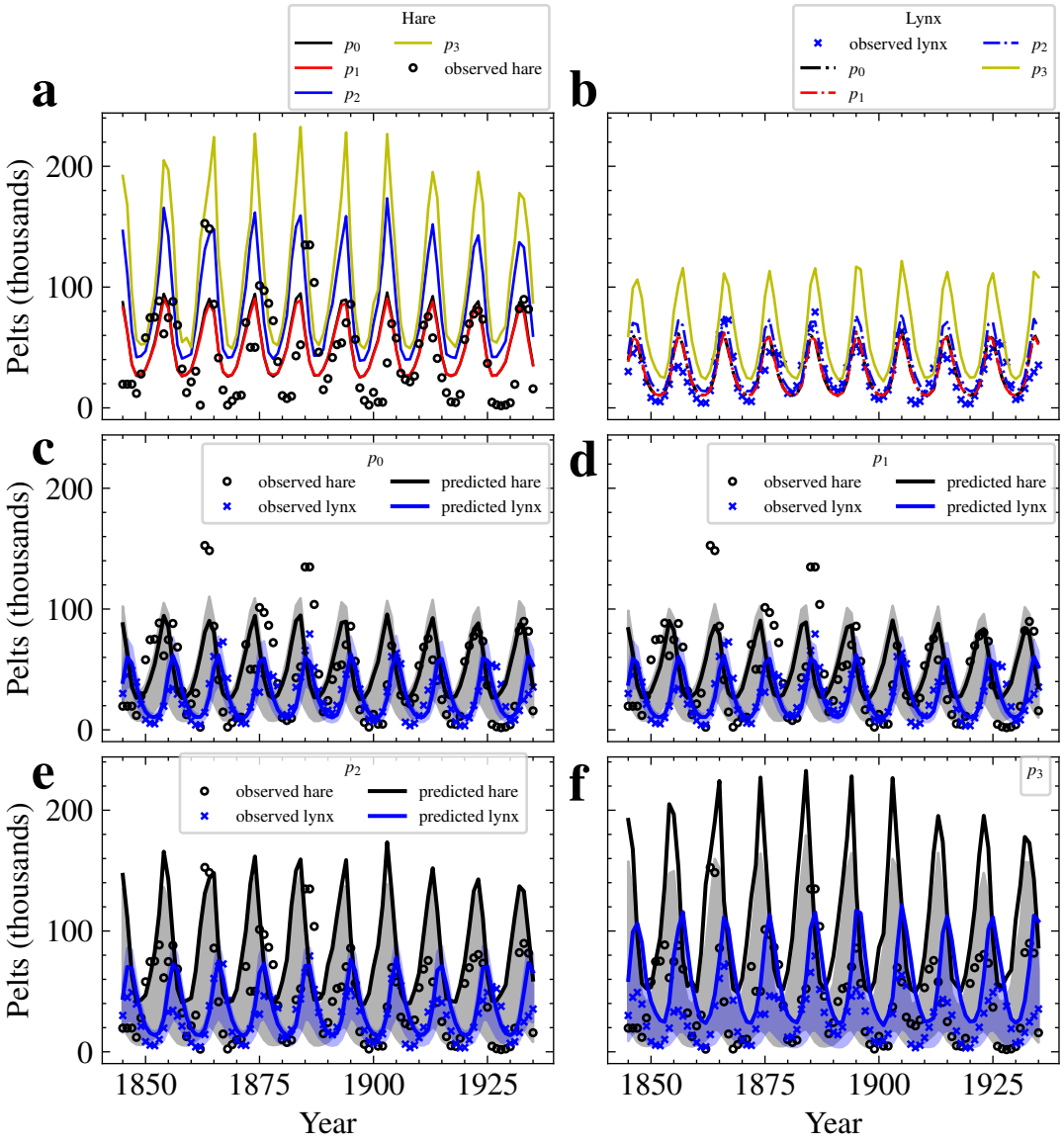


Figure 3.2: Graphical posterior predictive check for (a) Hare and (b) Lynx. The prior p_3 has a noticeable visual impact compared to p_0 and p_1. (c-f) Posterior predictive check for each prior with 25% and 75% quantiles.

evident when looking at the posterior estimates(Table 3.3) as the initial number of hares $\hat{\mu}$ for the baseline prior is further away from the other priors. The marginal sWIM is below one for the initial number of lynxs (ν). This is consistent with the posterior estimates Table 3.3 where the difference between the baseline prior and the others is very small. Again, this is not evident from the posterior predictive plots Fig. 3.2(b), where p_3 is visually different from the other priors.

Table 3.3: Posterior mean estimates of the different models for the Lotka-Volterra model.

parameter	p_0	p_1	p_2	p_3
	Baseline			
α	0.825	0.462	0.557	0.633
γ	0.037	0.019	0.023	0.028
β	0.404	1.038	0.894	0.810
δ	0.013	0.036	0.030	0.028
$\hat{u}$	16.427	48.662	53.451	53.378
$\hat{v}$	27.818	33.392	30.843	25.689
σ_u	1.054	1.042	1.413	1.597
σ_v	0.775	0.519	0.866	1.331

Table 3.4: Wasserstein-2 distances between prior p_i and posterior P_i distributions in the Lotka-Wolterra model, $i = 0, 1, 2, 3$. The values in bold are the WIM between the baseline posterior and the three other posteriors.

Wasserstein 2-distance						
	prior			posterior		
Baseline	p_1	p_2	p_3	P_1	P_2	P_3
p_0	36.266	26.975	25.124	28.864	40.504	39.40
P_0	49.647	51.257	49.320	**34.137**	**39.973**	**40.674**

Table 3.5: Prior scaled WIM between the baseline posterior and the three other posteriors in the Lotka-Volterra model.

	posterior		
Baseline posterior	P_1	P_2	P_3
P_0	0.941	1.482	1.619

Table 3.6: Marginal prior scaled WIM between the baseline posterior and the three other posteriors in the Lotka-Volterra model, only for parameters whose priors change across models.

	posterior		
parameter	P_1	P_2	P_3
$\hat{\mu}$	2.457	4.515	4.507
$\hat{\nu}$	0.415	0.646	0.812
σ_u	-	0.052	0.250
σ_v	-	0.013	0.275

3.3.2 SEIR model

The Bayesian approach is widely used for parameter estimation in epidemiology. We illustrate our prior assessment technique on the Susceptible-Exposed-Infectious-Recovered (SEIR) model, a commonly used epidemiological dynamical model. The COVID-19 pandemic in Luxembourg has been analysed in a Bayesian setting in Kemp et al. (2021) considering various control measures such as quarantine. Here, we calibrate the SEIR model for the first wave of the COVID-19 pandemic in Luxembourg using publicly available data found in (Mathieu et al., 2020). The first wave lasted from February to mid June 2020, a time frame that also similar studies determined as first wave (Grinsztajn et al., 2021). Compared to the original study (Kemp et al., 2021), we additionally perform posterior predictive checks and prior impact assessment using the proposed WIM and sWIM.

The SEIR model is defined by the following system of ODEs

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\eta S\frac{I}{N}, \quad S(t=0) = \hat{S} = \hat{N} - \hat{I} - \hat{E} - \hat{R}, \tag{3.13a}$$

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \eta S\frac{I}{N} - \sigma E, \quad E(t=0) = \hat{E}, \tag{3.13b}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \sigma E - \rho I, \quad I(t=0) = \hat{I}, \tag{3.13c}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \rho I, \quad R(t=0) = \hat{R} = 0, \tag{3.13d}$$

$$I_R = \lambda\left(\frac{\mathrm{d}E}{\mathrm{d}t} + \frac{\mathrm{d}I}{\mathrm{d}t}\right), \tag{3.13e}$$

$$N = S + E + I + R, \tag{3.13f}$$

where S is the number of susceptible individuals, that is, people not immune to COVID-19, E the number of exposed individuals, meaning people who have been infected but are themselves not yet infectious, I the number of infected and infectious individuals, R the number of recovered or deceased individuals, and I_R the rate of infection. N is the total number of individuals at any time which we assume to be constant.

The model has seven parameters, namely the transmission rate $\eta > 0$, the reciprocal of the incubation rate $\sigma > 0$, the recovery rate $\rho > 0$, the initial value for the number of infected individuals $\hat{I}$, the initial value for the number of exposed individuals $\hat{E}$, and a multiplicative correction for the rate of infection denoted by $\lambda \in (0, 1]$. This factor corrects for under-reporting and is the approach employed in (Grinsztajn et al., 2021).

We are interested in calibrating the model against the reported daily rate of infection I_R.

To that end, we employ a count distribution, more precisely the negative binomial NB with dispersion parameter $\phi > 0$:

$$y_i \mid \theta \sim \text{NB}([G_{I_R}(\theta)]_i, \phi), \quad i = 1, \ldots, n,$$

where $G_{I_R} : \mathbb{R}^p \to \mathbb{R}^n$ requires the solution of Eq. (3.13e) and its evaluation at n daily time points in the study period. We approximate I_R in Eq. (3.13e) with a first-order backwards finite differencing of Eqs. (3.13a) and (3.13b). Note that because $I_R \in \mathbb{R}$ we employ a non-standard negative binomial parametrization in TensorFlow Probability (Dillon et al., 2017) which naturally extends to the real numbers.

For the prior impact assessment, we use five sets of priors where the first set of priors is the baseline prior to which other priors are compared. The baseline prior p_0 and is similar to the non-informative priors used in (Grinsztajn et al., 2021; Moore et al., 2022) except for the moments of the distributions:

$$\begin{aligned}
\eta &\sim \text{TN}(2, 1), \\
\rho &\sim \text{TN}(0.4, 0.5), \\
\sigma &\sim \text{TN}(0.4, 0.5), \\
\hat{I} &\sim \text{TN}(0, 1), \\
\hat{E} &\sim \text{TN}(0, 1), \\
\lambda &\sim \text{Beta}(1, 2), \\
\phi^{-1} &\sim \text{Exponential}(5).
\end{aligned}$$

These priors are truncated normal distributions for most of the parameters which need to be positive, or zero in the case of $\hat{I}$ and $\hat{E}$.

The remaining five sets of priors shown in Table 3.7 where only the dispersion parameter ϕ varies across the priors. This is because overdispersion is usually the issue when modelling count data. Hence, it is important to see how overdispersion impacts inference. For the fifth set of priors not shown in Table 3.7, the dispersion parameter $\phi^{-1} \sim \text{Exponential}(150)$ is such that we can make a statement when the prior has a high impact compared to the baseline prior. It is noteworthy that we also choose a Gamma distribution as prior p_3, of which the Exponential is a special case.

The marginal posterior distributions for the baseline prior p_0 are shown in Fig. 3.3. The

posterior estimates for $\hat{I}$, $\hat{E}$ and ϕ have a higher standard deviation as shown by the density plots on the diagonal. Also, there is a noticeable correlation in the parameter pairs (λ, σ) and $(\lambda, \hat{I})$. The results of the posterior parameter estimates can be found in Table 3.8. The posterior mean estimates of ϕ for the priors p_1 and p_4 lie below one. For the Gamma prior p_3, the posterior estimate of the dispersion parameter is closer to that of the baseline prior p_0.

Fig. 3.4 shows posterior predictive checks for each prior with 25% and 75% confidence bands. The Gamma prior p_3 here resembles the most to the baseline prior p_0 but with a lower overdispersion parameter as shown in Table 3.8. This indicates that the gamma prior might be an alternative to the exponential prior on the overdispersion parameter. When the gamma prior is applied, the prediction intervals are narrower and enclose more observations than the baseline prior. In the case of the exponential prior, the mean value increases with the overdispersion parameter. The task is to address overdispersion without affecting the mean. Additionally, it is worth noting that the gamma distribution becomes the exponential distribution when the shape parameter equals 1. The mean predicted daily number of cases is quite close to the observed number for all priors except for p_4. For this prior, the 25% and 75% quantiles are also the widest compared to the others (see Fig. 3.3c and f).

Table 3.7: Priors used for the SEIR model. Only the prior on the dispersion parameter is different since overdispersion is usually the main modelling concern for count data. We choose the baseline prior p_0 like in other studies (Moore et al., 2022; Grinsztajn et al., 2021). A fifth prior not shown was also included.

p_0 (Baseline prior)	p_1	p_2	p_3
$\eta \sim \text{TN}(2, 1)$	$\eta \sim \text{TN}(2, 1)$	$\eta \sim \text{TN}(2, 1)$	$\eta \sim \text{TN}(2, 1)$
$\rho \sim \text{TN}(0.6, 0.5)$	$\rho \sim \text{TN}(0.6, 0.5)$	$\rho \sim \text{TN}(0.6, 0.5)$	$\rho \sim \text{TN}(0.6, 0.5)$
$\sigma \sim \text{TN}(0.4, 0.5)$	$\sigma \sim \text{TN}(0.4, 0.5)$	$\sigma \sim \text{TN}(0.4, 0.5)$	$\sigma \sim \text{TN}(0.4, 0.5)$
$\hat{I}, \hat{E} \overset{\text{iid}}{\sim} \text{TN}(0, 1)$	$\hat{I}, \hat{E} \overset{\text{iid}}{\sim} \text{TN}(0, 1)$	$\hat{I}, \hat{E} \overset{\text{iid}}{\sim} \text{TN}(0, 1)$	$\hat{I}, \hat{E} \overset{\text{iid}}{\sim} \text{TN}(0, 1)$
$\lambda \sim \text{Beta}(1, 2)$	$\lambda \sim \text{Beta}(1, 2)$	$\lambda \sim \text{Beta}(1, 2)$	$\lambda \sim \text{Beta}(1, 2)$
$\phi^{-1} \sim \text{Exponential}(5)$	$\phi^{-1} \sim \text{Exponential}(42)$	$\phi^{-1} \sim \text{Exponential}(1)$	$\phi^{-1} \sim \text{Gamma}(16, 16)$

Let us now discuss the findings based on WIM and sWIM. The results for the WIM are provided in Table 3.9. The WIM is highest for the prior p_4, which is totally in line with the posterior predictive check where the plot is clearly distinguishable from the other plots and hence in particular from that corresponding to p_0. The same holds for p_1, though with a smaller WIM and this is consistent with a lower peak in Fig. 3.4a. The sWIM is shown in Table 3.10. We can see that the sWIM is below 1 for p_2 and p_3 while above 1 for p_1 and p_4. This example illustrates that the gamma prior (p_3) has similar impact on the posterior as the baseline prior (p_0). For p_1

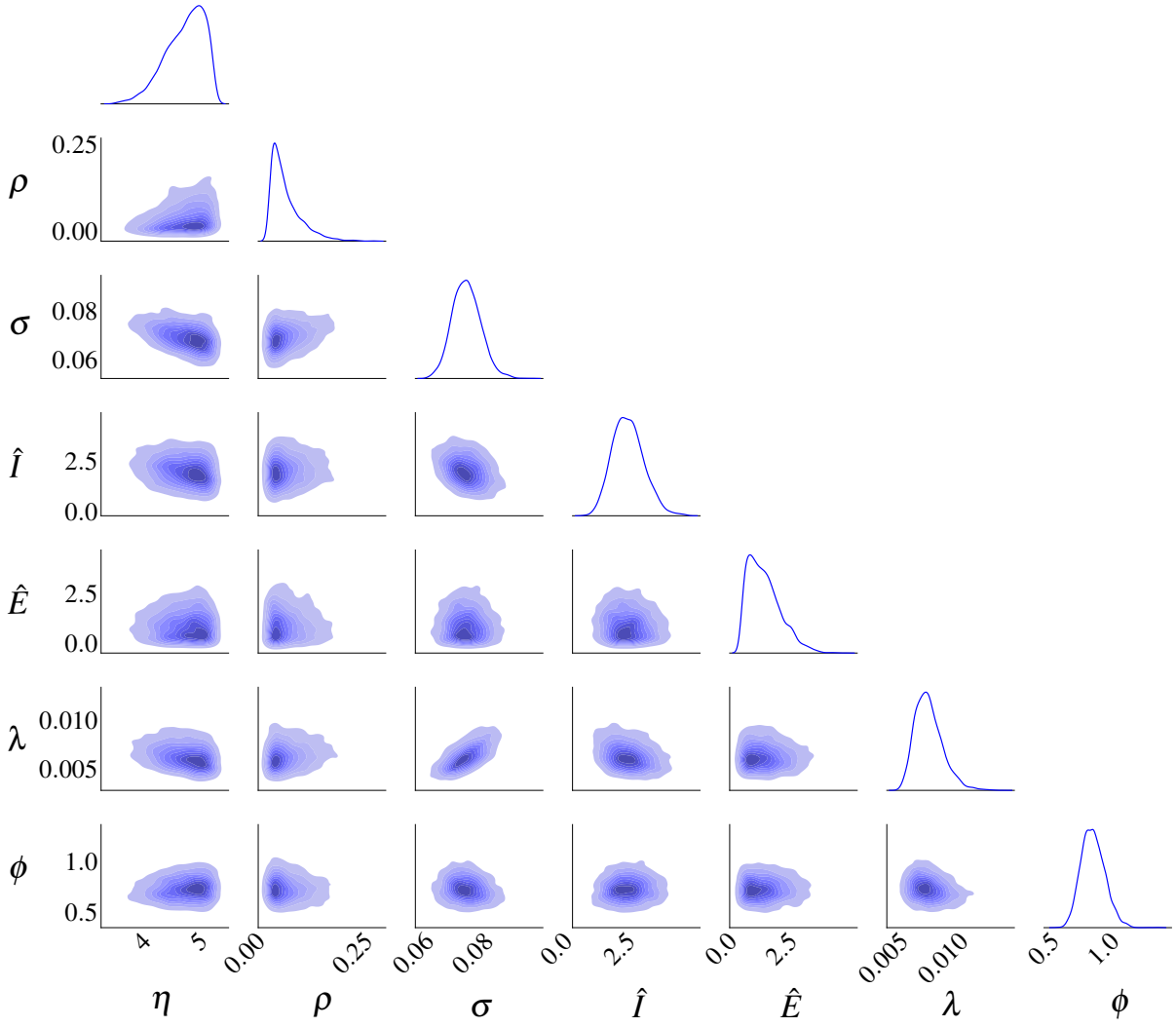


Figure 3.3: Plot of posterior distributions associated to p_0 in the SEIR model, with posterior marginal distributions on the diagonal and bivariate distributions for outside the diagonal. There is a high correlation between the parameters λ and σ.

Table 3.8: Posterior mean estimates of the different models for the SEIR model.

parameter	p_0	p_1	p_2	p_3	p_4
η	4.689	4.497	4.729	4.678	4.161
ρ	0.024	0.036	0.022	0.025	0.059
σ	0.066	0.069	0.066	0.066	0.076
$\hat{I}$	2.255	2.000	2.294	2.235	1.701
$\hat{E}$	0.891	0.908	0.901	0.886	0.881
λ	0.005	0.006	0.005	0.005	0.009
ϕ	1.856	0.734	2.426	1.674	0.315

with sWIM of 1.276, the parameter estimates are still close to those of the baseline posterior, although the posterior predictive check shows higher predictions than observed. In the case of p_4 with sWIM of 1.762, the parameter estimates are further from the baseline posterior estimates, and the posterior predictive check shows higher predictions than observed.

Table 3.9: Wasserstein-2 distances between prior p_i and posterior P_i distributions in the SEIR model, $i = 0, 1, 2, 3$. The values in bold are the WIM between the baseline posterior and the four other posteriors.

Wasserstein-2 distance								
	prior				posterior			
Baseline	p_1	p_2	p_3	p_4	P_1	P_2	P_3	P_4
p_0	1.1055	1.650	1.401	1.109	3.299	4.240	3.811	1.109
P_0	3.991	3.747	3.644	4.000	**1.410**	**0.925**	**0.688**	**1.954**

The marginal sWIM was also computed for ϕ, and the results are in Table 3.11. The marginal sWIM values are less than 1 for p_2 and p_3, which ties to parameter estimates and predictions closer to baseline.

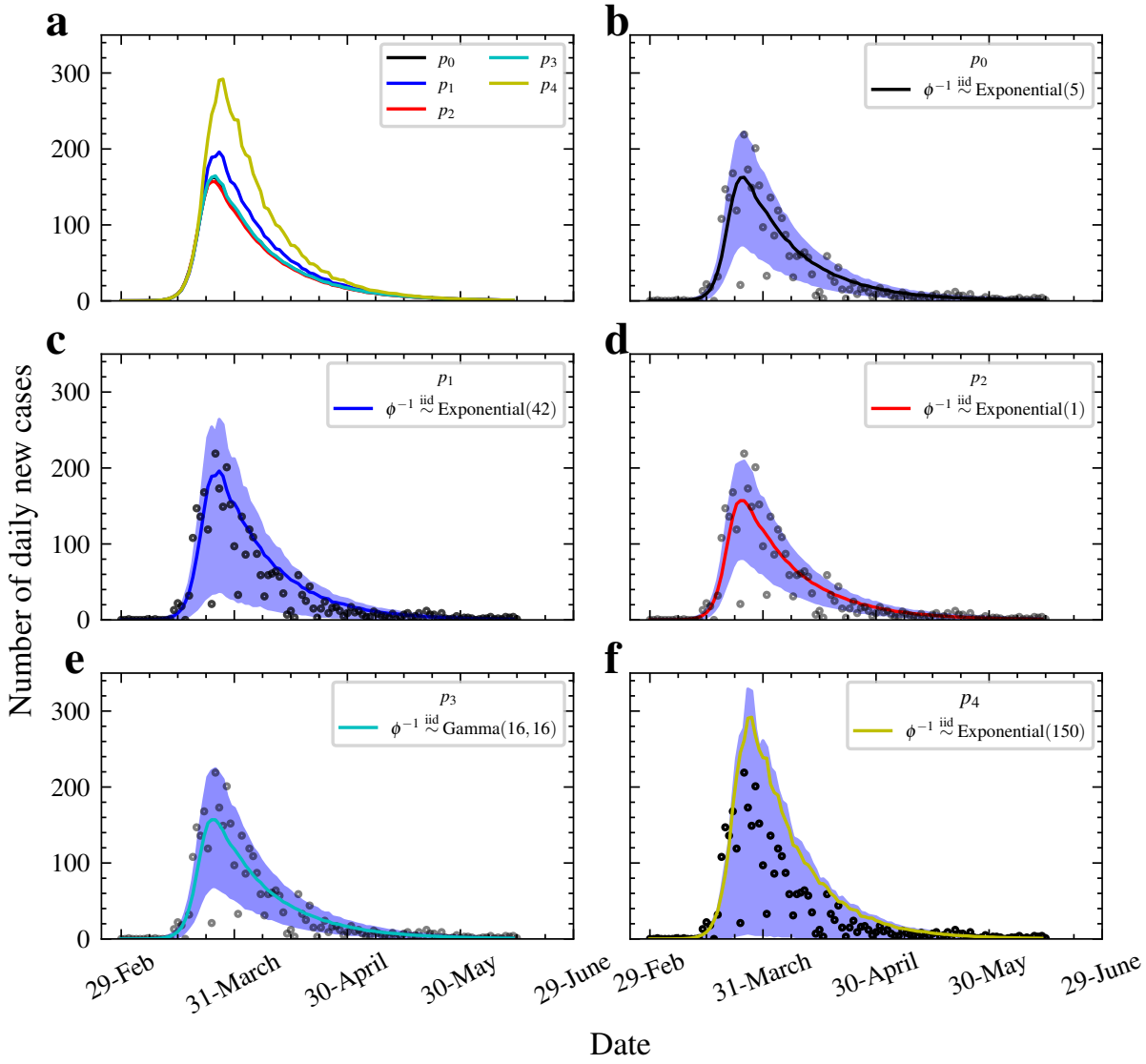


Figure 3.4: (a) Graphical posterior predictive check for all priors in the SEIR model, and (b-f) posterior predictive check for each prior with 25% and 75% quantiles. The Gamma(16,16) prior seems to be a better option since the posterior has less variability compared to the other. Moreover, most of the observed counts are in the 25% to 75% prediction bands unlike for other posteriors considered where the highest and lowest counts are outside, or the bands are wider like for Exponential(42). The prior p_4 has the largest predicted values and the predictions are further away from the observed values compared to the other priors.

Table 3.10: Prior scaled WIM between the baseline posterior and the four other posteriors in the SEIR model.

sWIM				
	posterior			
Baseline posterior	P_1	P_2	P_3	P_4
P_0	1.276	0.561	0.491	1.762

Table 3.11: Marginal prior scaled WIM between the baseline posterior and the four other posteriors in the SEIR model, only for parameters whose priors change across models.

sWIM				
	posterior			
parameter	P_1	P_2	P_3	P_4
ϕ	6.245	0.855	0.324	7.866

3.4 Conclusions

This study employs computational optimal transport to quantify the difference in posterior inference between different priors for dynamic systems modelled by ODEs. Using the Sinkhorn algorithm from computational optimal transport, we can rapidly compute the Wasserstein distance for multiparameter systems. Building on this, we have extended the WIM as prior impact assessment tool to ordinary differential equations. We have also introduced the sWIM, an interpretable impact measure. Like the WIM, it can be quickly calculated. When sWIM < 1, the prior in question has no greater impact than the baseline prior. When sWIM > 1, the prior in question has a higher impact than a baseline prior. It is also insightful to compute the marginal sWIM, meaning the sWIM for each parameter instead of the joint parameters. Our approach has been exemplified with the Lotka–Volterra predator-prey model and the SEIR for COVID-19. For both examples, we used real-world data. The results show that the difference in posterior mean estimates is close to zero when sWIM < 1. In addition, graphical posterior predictive checks show that predictions are closer to the baseline when sWIM < 1 and further when sWIM > 1. In future research, our goal is to decide by means of an extensive simulation study at what values of the sWIM one should label a prior as high or low impact relative to a baseline prior.

Code and data availability

The full code to produce the results in this article is available at D. N. Mingo & Hale (2024) and GitHub at BayesianODE-PriorImpactAssessment.

Credit contributor roles

DNM: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing - review & editing. JSH: Conceptualization, Formal analysis, Funding acquisition, Resources, Methodology, Project administration, Software, Supervision, Validation, Writing - original draft, Writing - review & editing. CL: Conceptualization, Funding acquisition, Formal analysis, Methodology, Supervision, Writing - review & editing.

Competing interests

The authors declare no competing interests.

Acknowledgements

This work was funded under the Luxembourg National Research Fund under the PRIDE programme (PRIDE17/12252781).

The experiments were carried out using the HPC facilities of the University of Luxembourg (Varrette et al., 2022) – see `https://hpc.uni.lu`.

Chapter 4

Insights into the prior to posterior transition through Wasserstein distances and the power posterior

The content of this chapter is based on the following paper in preparation.

Mingo, D. N., and Hale, J. S. (2024). Insights into the prior to posterior transition through Wasserstein distances and the power posterior[in preparation]

Abstract

Several studies have analysed how priors impact posteriors. However, the evolution from prior to posterior is not well understood. This study investigates how priors transition to posteriors and provides practical insights. This is done by calculating the Wasserstein distance between the priors and the power posteriors and by introducing the concept of saturated sample size. The Wasserstein distance is a measure of the difference between two probability distributions. Power posteriors are obtained by raising the likelihood function to values between zero and one, creating a continuous path from the prior to the full posterior. The saturated sample size is the minimal number of observations that contain the same amount of information as the entire dataset. The saturated sample size is important when obtaining larger sample sizes is challenging, such as in clinical trials for rare diseases or when individuals are reluctant to participate in surveys. We have demonstrated with two examples how using the concept of saturated sample size allows us to decrease the sample size from 1000 to 300 and from 100 to 70.

4.1 Introduction

Priors are a crucial element of Bayesian analysis, and the choices made can significantly impact parameter inference. Various methods have been developed to assess the impact of priors. However, there is little understanding of how priors evolve to the posterior. Power posteriors form a bridge between the prior and the full posterior (Meng & Wong, 1996; Gelman & Meng, 1998). This bridge is constructed by raising the likelihood to values between zero and 1. Distances such as the Wasserstein distance can be used to quantify how these power posteriors are similar to the prior. This study employs power posteriors and Wasserstein distances to gain insights into prior to posterior transitions. Power posteriors have been used in many applications, such as calculating the marginal likelihood (Friel & Pettitt, 2008) and calculating the objective BF (O'Hagan, 1997), where they are referred to as fractional posteriors. They have also been applied to sample across multimodal distributions and are used to set up power priors (Fouskakis et al., 2015; Ibrahim & Chen, 2000) for robust Bayesian inference (Miller & Dunson, 2019). For conjugate cases, we derive and make use of analytic power posteriors. We use graphs to illustrate how the priors transition to the posteriors.

The sample size is critical for inference. However, having large sample sizes for some studies

might be difficult. Fortunately, some sample sizes can give the same amount of information as large sample sizes. Thus, linking sample sizes, power posteriors, and saturated sample sizes is essential.

4.1.1 Contributions

The main contributions of this chapter are as follows. Using the concept of the power posterior we construct a continuous family of posteriors between prior and posterior by raising the usual likelihood to a power posterior constant $\gamma \in [0, 1]$. To give a visual intuition of this transition, posterior density plots of the power posteriors living between a prior and the standard posterior for a parameter can be made. From the definition of the power posterior we derive extensions of three standard conjugate results, namely the Normal-Normal, Normal-Inverse-Gamma and Poisson-Gamma cases. These conjugate results show that for an observed dataset of size n the product γn as the effective sample size of the observed data, that is, γ can be used to reduce the sample size of the data from n to $n\gamma$.

With the definition of the power posterior in hand we then observe the evolution of the Wasserstein distance between power posterior measures μ_γ and the fixed prior measure μ_0. These plots typically show a sudden change in the Wasserstein distance as γ is increased just beyond zero (the prior is moved a large distance by a small effective sample size of the data). As γ is increased at some point the distance 'saturates' and increasing γ further (i.e. increasing the effective sample size) does not lead to the posterior moving significantly further away from the prior. This saturation region gives us an indication of a saturated sample size, that is, at which adding more data into the problem (implicitly via the parameter γ) does not lead to an increase in the distance between the power posterior and the prior. The saturated sample size is the total sample size n scaled by the value of γ at which the Wasserstein distance between the prior and the standard posterior appears to stabilise. We demonstrate through simulations that subsampling is equivalent to power posteriors, where the power γ serves as a weight for the data, resulting in a reduced sample size.

In addition to looking at an prior to posterior transition for a single fixed prior, we also look at the evolution of the Wasserstein distance under different prior assumptions with the findings illustrated in graphs.

4.2 Methodology

This section covers the concepts and methodologies used in this study, such as power posteriors, Wasserstein distance calculations, subsampling, and numerical methods.

4.2.1 Power posterior

We briefly recall the mathematical concept of the power posterior (sometimes also called the fractional posterior) which is central to this paper. Let x be some observed data and $p(x \,|\, \theta)$ be the likelihood function of a model that depends on a set of parameters θ with associated prior $p(\theta)$. Moreover, let γ be an additional *power posterior constant* in the interval $[0, 1]$. Then the power posterior (Friel & Pettitt, 2008) can be defined as

$$p(\theta \,|\, x, \gamma) := \frac{p(x \,|\, \theta)^{\gamma} p(\theta)}{z}, \quad \gamma \in [0, 1], \tag{4.1}$$

where $z := p(x)$ is a normalising constant called the marginal likelihood. It is straightforward to see that for $\gamma = 0$ we recover the usual Bayesian prior $p(\theta)$ and for $\gamma = 1$ we recover the standard Bayesian posterior $p(\theta \,|y)$. We emphasise that for intermediate values $\gamma \in (0, 1)$ the resulting set of power posteriors forms a continuous transition between the standard prior and posterior distribution. The likelihood $p(x|\theta)$ reflects how likely it is for a parameter to generate the data. Taking the log of Eq. (4.1) we can see from $\gamma \log p(x|\theta)$ that γ controls the relative weight of the log-likelihood (and consequently the data) with respect to the log-prior.

4.2.2 Wasserstein distance

Various measures can be used to quantify the discrepancies between distributions. Some of these measures involve computing the KL and chi-square divergence. However, these divergences (Ali & Silvey, 1966) are not symmetric, and even when symmetric versions exist, other challenges arise. For instance, the KL divergence is undefined if the intersection of the support of the two distributions in question is a null set. In contrast, Wasserstein distance is symmetric, can be computed between discrete and continuous probability distributions, considers the geometry of the parameter space (Panaretos & Zemel, 2019), and is a well-defined metric. Therefore, we employ the Wasserstein distance. The definition of the p-Wasserstein distance (W_p) between

two probability measures μ, ν defined on the space $\mathcal{X}$ is

$$W_p(\mu, \nu) = \inf_{\pi \in U(\mu,\nu)} \left(\int_{\mathcal{X}\times\mathcal{X}} ||x - y||^p d\pi(x, y) \right)^{1/p}, \quad p \geq 1, \tag{4.2}$$

where $U(\mu, \nu)$ is the set of joint probability measures on $\mathcal{X} \times \mathcal{X}$ (Villani, 2009), x and y are points in the space X. The Wasserstein distance has a natural interpretation as the minimum amount of work required to reconfigure the mass of one distribution into another. Having introduced power posteriors and the Wasserstein distance, we now formally define subsampling and the saturated sample size

4.2.3 Subsampling

Subsampling refers to the method of drawing a random sample from a dataset to analyse it as a representative of the complete dataset (Drineas et al., 2006; Yao & Wang, 2021; Ma et al., 2014). Consider $\{x_i\}_{i=1}^n$ are n independent and identically distributed observations. We can draw a subsample of size m from the entire sample of size n with sampling probabilities $\{\pi_i\}_{i=1}^n$ on each observation. The observations have equal sampling probabilities. The motivation for subsampling is to avoid the computational expense of analysing the entire dataset.

Subsampling is similar to using power posteriors. When working with large datasets, using power posteriors with a power value close to 1 can be a practical strategy to manage computational complexity and reduce the impact of model misspecification. To demonstrate this, we calculated the expectation of the Wasserstein distance between the priors and the standard posterior across multiple subsamples of size $m = \gamma n$, where m is the subsample size, n is the size of the entire dataset, and γ varies for each subsample. We also calculated the Wasserstein distances between the priors and their corresponding power posteriors. This enables us to compare the Wasserstein distance between the prior and the standard posterior with respect to subsampling and the Wasserstein distance between the prior and the power posteriors.

4.2.4 Saturated sample size

The saturated sample size m_{sat} is the total sample size n scaled by the value of γ at which the Wasserstein distance between the prior and the standard posterior appears to stabilise. It is defined as

$$m_{\text{sat}} = \gamma_{W_t} \cdot n, \tag{4.3}$$

where γ_{W_t} is the value of γ at the saturated Wasserstein distance W_t defined as

$$W_t = \text{CrI} \cdot W_{\mu_0,\gamma_1},$$

where CrI a threshold which can be 95% or 99% , W_{μ_0,γ_1} is the Wasserstein distance between the prior and the standard posterior. When $W_t = 1$, Eq. (4.3) becomes

$$m_{\text{sat}} = m = \gamma n, \tag{4.4}$$

where m is the subsample size, n is the size of the entire dataset, and $\gamma \in [0, 1]$.

4.2.5 Posterior measures

In this section, we introduce several measures that will be important in the results section. The first measure is the prior distribution, μ_0, representing our initial beliefs or assumptions before analysing any data. Next, we have the power posterior, μ_γ, which is a variation of the standard posterior that raises the likelihood to a constant γ, ranging from 0 to 1, to adjust the influence of the data on inference. Another key measure is the squared 2-Wasserstein distance between the power posterior and the prior distribution $W_2^2(\mu_\gamma, \mu_0)$. Moreover, we define the expectation of this squared 2-Wasserstein distance between a standard posterior μ_1 and the prior μ_0 as the average of these distances across multiple subsamples from the dataset. This expectation is calculated as

$$E[W_2^2(\mu_1, \mu_0)] = \frac{1}{S} \sum_{i=1}^{S} W_2^2(\mu_1, \mu_0), \tag{4.5}$$

where S represents the number of subsamples, each of size m, drawn from the full dataset of size n.

4.2.6 Numerical methods

We give a brief overview of the numerical methods used to generate the results. For full details see the references provided below.

Samples from the power posteriors and priors are generated using MCMC, specifically the NUTS (Hoffman & Gelman, 2014). We use the version implemented within TFP (Dillon et al., 2017) running on the JAX backend (Bradbury et al., 2018).

We remark that exploring power posteriors within TFP is straightforward and only requires the programmatic definition of the log power-posterior. This involves scaling the provided log-likelihood evaluation by γ and then adding the log-prior terms. This can be achieved in a few lines of code and the resulting function passed to any of the MCMC methods built-in to TFP.

The skew-normal distribution is not implemented in TFP, so we wrote a custom distribution using Distrax (DeepMind et al., 2020) following the numerical methodology described in (Ghorbanzadeh et al., 2014).

To calculate Wasserstein-1 distance for the one-dimensional problems we use the Vallender formula (Vallender, 1974) which relates the 1-Wasserstein distance between two probability measures μ_1 and μ_2 on $\mathbb{R}$ with cumulative distribution functions $F_1(x)$ and $F_2(x)$, respectively

$$W_1(\mu_1, \mu_2) = \int_{\mathbb{R}} |F_1(x) - F_2(x)| \, \mathrm{d}x. \tag{4.6}$$

We approximate the Vallender formula using the dedicated function available in SciPy (Virtanen et al., 2020).

For the multi-dimensional case we cannot use to the Vallender formula to compute the Wasserstein distance. Instead we use recent advances in computational optimal transport to estimate the Wasserstein distance. Specifically, we use entropic regularised optimal transport proposed in (Cuturi, 2013) and its implementation within the OTT library (Cuturi et al., 2022).

All code and data used to produce the results in this paper will be available as supplementary material upon journal submission.

4.3 Summary of methodology

This section gives a concise overview of the methodology. We employ conjugate cases where the power posteriors are available in closed. A conjugate case occurs when the prior and the posterior distributions are of the same family. We formally introduce the concept of saturated sample size, which links subsampling to power posteriors. The connection is demonstrated by subsampling with different sample sizes.

4.3.1 Evolution of the Wasserstein distances

1. Generate synthetic data from the models for calibration.

2. Set up different priors and a likelihood function. Obtain the power posteriors. For conjugate cases, the power posteriors are available in closed form.

3. Compute the Wasserstein distances between the priors and their corresponding power posteriors.

4.3.2 Subsampling

1. Set up a sequence of γ values from 0 to 1. For each value of γ, compute the Wasserstein distances between the priors and the power posteriors using the entire dataset.

2. For each value of γ, draw S subsamples, each of size m from the entire dataset, where

$$m = \gamma n,$$

   with $\gamma \in [0, 1]$ representing the power, and n is the sample size of the entire dataset.

3. For each subsample, compute the Wasserstein distances between the priors and the respective standard posteriors.

4. Compute the expectation of the Wasserstein distances across the S subsamples.

5. Also, take note of the saturated sample size.

The saturated sample size is the number of observations needed to achieve a value of γ at which the Wasserstein distance between the power posterior and the prior reaches 95% or 99% of the Wasserstein distance between the full posterior ($\gamma = 1$) and the prior($\gamma = 0$). In other words, the saturated sample size is the number of observations required to reach a critical value of γ.

4.4 Results

This section presents the results of the transition of various priors to their corresponding posteriors. These results are based on experiments with power posteriors for conjugate cases and a skew-normal distribution. Also, we illustrate that subsampling yields results equivalent to those obtained through power posteriors.

4.4.1 Normal-normal conjugate case with unknown mean

Following standard arguments, it is possible to show that the distribution of the power posterior for the following Bayesian model with dataset $x = (x_1, \ldots, x_n)$

$$x_1, \ldots, x_n \overset{\text{iid}}{\sim} \mathcal{N}(m, \sigma^2), \tag{4.7a}$$

$$m \sim \mathcal{N}(m_0, \sigma_0^2), \tag{4.7b}$$

is normally distributed (see Section A.1, Theorem 1 for proof)

$$m \mid \bar{x}, \gamma \sim \mathcal{N}\left(\left(\frac{1}{\sigma_0^2} + \frac{\gamma n}{\sigma^2}\right)^{-1}\left(\frac{m_0}{\sigma_0^2} + \frac{\gamma n \bar{x}}{\sigma^2}\right), \left(\frac{1}{\sigma_0^2} + \frac{\gamma n}{\sigma^2}\right)^{-1}\right),$$

where $\bar{x} = (\sum x_i)/n$ is the sample mean. We remark that, as expected, the prior is recovered when $\gamma = 0$ and the classic normal-normal with unknown mean conjugate result is recovered when $\gamma = 1$. The role of γ in this context is to reduce the contribution of each element of the dataset through the likelihood. More specifically for the normal-normal case, the effective dataset size is reduced from the standard posterior ($\gamma = 1$) from n to $n\gamma$ in the power posterior. Note however, regardless of the value of $\gamma > 0$, the entire dataset x is still used in the update from prior to the power posterior.

The 2-Wasserstein metric ($p = 2$) for two non-degenerate normal measures μ_1 and μ_2 on $\mathbb{R}$ with means $m_1, m_2 \in \mathbb{R}$ and variances $\sigma_1^2, \sigma_2^2 \in \mathbb{R}_{>0} := \{x \in \mathbb{R} \mid x > 0\}$, respectively, can be found in closed form as

$$W_2(\mu_1, \mu_2)^2 = (m_1 - m_2)^2 + \sigma_1^2 + \sigma_2^2 - 2\big(\sigma_2 \sigma_1^2 \sigma_2\big)^{1/2}.$$

Consequently in this case there is no need to resort to approximate numerical computations to compute the Wasserstein metric. In closed form the Wasserstein metric between the prior measure μ_0 and the measure induced by the power posterior μ_γ is

$$W_2(\mu_0, \mu_\gamma)^2 = \sigma_0^2 + \frac{\sigma^2 \sigma_0^2}{p} + \frac{\big(\gamma m_0 n \sigma_0^2 - \gamma n \sigma_0^2 \bar{x}\big)^2}{p^2} - \frac{2\sigma\sigma_0^2}{\sqrt{p}}$$

with $p = \gamma n \sigma_0^2 + \sigma^2$. Letting $\gamma = 0$ it can be verified that $W_2(\mu_0, \mu_0)^2 = 0$.

In the first experiment we generate a (small) dataset of size $n = 10$ from a normal distribution with zero mean and unit variance. We propose three prior choices (Table 4.1) by adjusting the

prior parameters m_0 and σ_0^2:

1. *Non-informative prior.* With this relatively flat prior we expect the posterior to be 'data prominent' and highly sensitive to the inclusion of information via the likelihood, here controlled by the parameter γ.

2. *Informative 'correct' prior.* This prior expresses definite information about the parameter that coincidentally coincides with the true parameter.

3. *Informative 'incorrect' prior.* This prior expresses definite information about the parameter that does not coincide with the true parameter value.

Table 4.1: Priors and their corresponding distributions the Normal likelihood normal prior case.

Priors	Distribution
Non informative	$\mathcal{N}(0.0, 100)$
Informative 'correct'	$\mathcal{N}(0.0, 2)$
Informative 'incorrect'	$\mathcal{N}(-5.0, 2)$

In Fig. 4.1 we calculate the squared 2-Wasserstein distance between the posterior measure μ_γ for varying values of γ and under the three prior assumptions just described. All three distances appear to be monotonically increasing in γ. The distance between the posterior and prior as $\gamma \to 1$ is largest ($\sim 10^2$) for the non-informative prior, and smallest (~ 1) for the informative 'correct' prior. For the non-informative prior as γ is increased from 0 we can see a very sudden increase in the metric, which supports the intuition that the posterior is 'data prominent' or 'data sensitive'. By comparison, the distance for the informative 'correct' prior evolves to its final distance more slowly, reflecting that the information contained in the prior does not strongly disagree with that in the data. The informative 'incorrect' prior sits in between the two extreme cases.

Given the interpretation of the parameter γn as an effective sample size in the normal-normal setting, we briefly contrast the power posterior approach with an approach based on using the standard posterior and varying the size n of the dataset.

To demonstrate the relationship between subsample size and the power posterior, we begin by generating 1000 samples from a standard normal distribution. We then draw 500 subsamples,

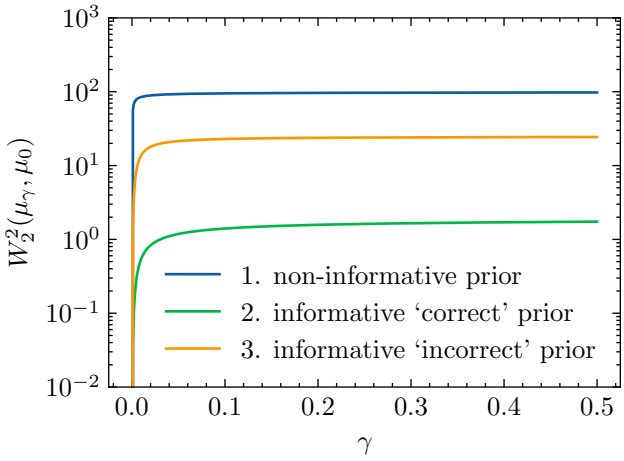


Figure 4.1: Normal-normal conjugate case; plot showing squared 2-Wasserstein metric between power posterior and prior against the power posterior parameter $\gamma \in [0, \frac{1}{2}]$ under the three different prior assumptions described in the text.

each of size $m = \gamma n$ where $n = 1000$ and γ varies from 0 to 1. For each subsample, we compute the squared 2-Wasserstein distance between the prior and the standard posterior, as well as the mean of these Wasserstein distances, using Eq. (4.5). The results, along with the corresponding $\pm 2\sigma$ interval around the mean, are presented in Fig. 4.2. Additionally, we calculate the squared 2-Wasserstein distance between the prior and the power posterior for various values of γ.

In addition, Fig. 4.2 illustrates the convergence of the Wasserstein distance 1) between the prior ($\gamma = 0$) and the power posteriors $\gamma \in (> 0, 1)$ and 2) between the prior and the standard posteriors ($\gamma = 1$) as γ increases. As γ approaches 1, this convergence exemplifies the increasing similarity between the power and standard posteriors with subsampling. Also, the uncertainty in the standard posterior estimates decreases as γ approaches 1, or in other words, as the subsample size increases. For higher values of γ, the power posteriors and the standard posteriors with subsampling will lead to similar results.

The saturated sample size is also shown in Table 4.2. We see that sample sizes can be reduced from 1000 to 30 or 300 depending on the chosen threshold without compromising the information provided by the data.

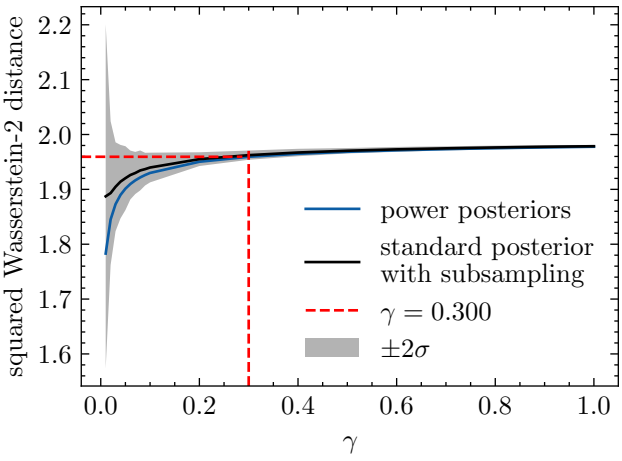


Figure 4.2: Normal-normal conjugate case with unknown mean. Wasserstein distance between the prior and power posteriors compared to the distance between the prior and standard posterior with subsampling. The dashed lines represent the Wasserstein distance and the parameter γ at the saturated sample size at a threshold of 99%.

Table 4.2: Saturated sample size and threshold for the Normal prior Normal likelihood case.

Wasserstein distance (W_t)	Threshold (CrI)	γ_{W_t}	Total sample size (n)	Saturated sample size (m_{sat})
1.874	95%	0.03	1000	30
1.961	99%	0.30	1000	300

4.4.2 Inverse-Gamma conjugate case with unknown variance

Following standard arguments, it is possible to show that the distribution of the power posterior for the following Bayesian model with dataset $x = (x_1, \ldots, x_n)$

$$x_1, \ldots, x_n \overset{\text{iid}}{\sim} \mathcal{N}(m, \sigma^2), \tag{4.8a}$$

$$\sigma^2 \sim \text{Inv-Gamma}(\alpha, \beta) \tag{4.8b}$$

with $\alpha, \beta \in \mathbb{R}_{>0}$, is distributed according to an inverse gamma distribution (see Section A.1, Theorem 2 for proof)

$$\sigma^2 \mid s^2, \gamma \sim \text{Inv-Gamma}\left(\alpha + \frac{\gamma n}{2}, \beta + \frac{\gamma n}{2} s^2\right).$$

where $s^2 = (\sum_i (x_i - \mu)^2)/n$ is the uncorrected sample variance. Again, setting $\gamma = 0$ recovers the prior, and setting $\gamma = 1$ recovers the usual conjugate result for the standard Bayesian posterior.

We explore how the Wasserstein distance between the prior and power posteriors changes. Firstly, we generate data of size 100 from a standard normal distribution. Then, an inverse gamma prior with different values of the shape and scale parameters is used, giving three sets of priors (Table 4.3): informative 'correct', non-informative prior, and informative 'incorrect' prior.

Table 4.3: Priors and their corresponding distributions.

Priors	Distribution
Non informative	Inv-Gamma$(1.0, 1.0)$
Informative 'incorrect'	Inv-Gamma$(3.0, 0.5)$
Informative 'correct'	Inv-Gamma$(3.0, \sqrt{2})$

The priors are used to fit three different models to the data with a normal likelihood. For each prior, the Wasserstein distance between the prior and the power posteriors is calculated. When using a non-informative prior, this distance increases with the value of γ, eventually levels off at a γ value between 0.1 and 0.2 (Fig. 4.3). In contrast, the distance decreases initially for informative priors but eventually stabilises. Nonetheless, as γ increases, all priors converge to the same distance due to incorporating more data.

4.4.3 Poisson-Gamma conjugate case with unknown rate

The distribution of the power posterior of the following Bayesian model with dataset $x = (x_1, \ldots, x_n)$

$$x_1, \ldots, x_n \overset{\text{iid}}{\sim} \text{Poisson}(\lambda), \tag{4.9a}$$

$$\lambda \sim \text{Gamma}(\alpha, \beta), \tag{4.9b}$$

where $\lambda \in \mathbb{R}_{>0}$ is the unknown rate parameter, and $\alpha, \beta \in \mathbb{R}_{>0}$ are prior parameters is given by (see Section A.1, Theorem 3 for proof)

$$\lambda \mid \bar{x} \sim \text{Gamma}\left(\alpha + \gamma n \bar{x}, \beta + \gamma n\right) \tag{4.10}$$

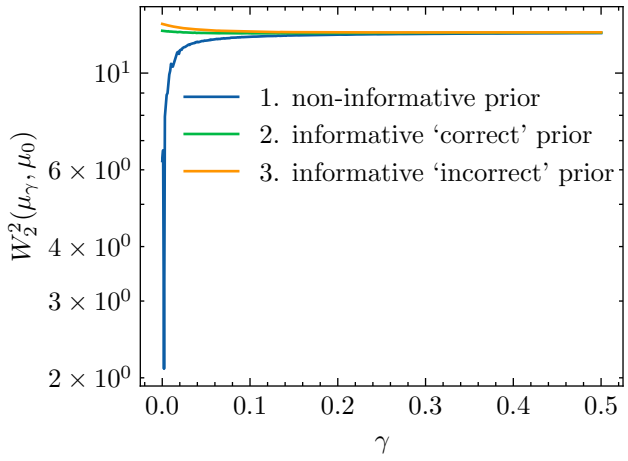


Figure 4.3: Inverse-Gamma conjugate case with unknown variance; plot showing squared 2-Wasserstein metric between power posterior and prior against the power posterior parameter $\gamma \in [0, \frac{1}{2}]$ under the three different prior assumptions described in Table 4.3.

We investigate the impact of three different sets of priors on the posterior distribution. We generate 100 observations from a Poisson distribution with a rate parameter 1.0. We employ three sets of priors Table 4.4 with a Poisson likelihood for the analysis.

Table 4.4: Priors and their corresponding distributions.

Priors	Distribution
Non informative	Gamma$(1.0, 1.0)$
Informative 'incorrect'	Gamma$(10.0, 1.0)$
Informative 'correct'	Gamma$(1.0, 0.2)$

We compute the squared Wasserstein distance between each prior and its corresponding power posterior. Our analysis indicates that as γ increases, the distance also increases for non-informative priors, while it decreases with increasing γ for informative but incorrect priors Fig. 4.4. Eventually, the distances stabilise across all priors. Informative priors converge to the same values, and an elbow point is observed for all priors, beyond which the distance levels off. For the Inverse Gamma prior, this stabilisation occurs just before reaching a value of 0.2.

To further study the relation between sampling and γ , we draw 50 subsamples, each of size m where $m = \gamma n$ with $n = 100$. We then compute the squared Wasserstein distance between the prior and each subsample for different values of γ and then the mean of the Wasserstein distance. Additionally, we calculate the distance between the priors and the power posteriors,

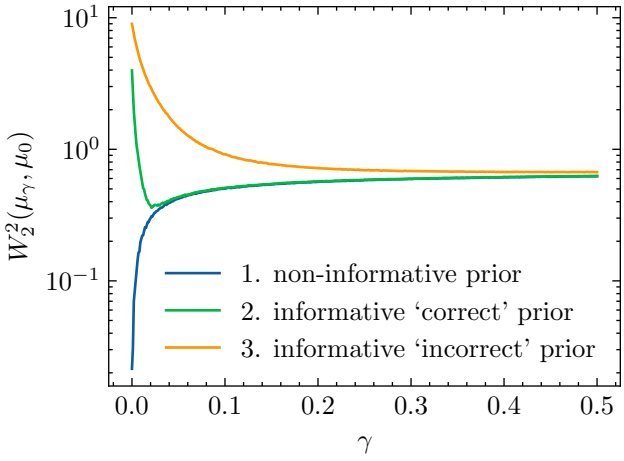


Figure 4.4: Poisson-Gamma conjugate case with unknown rate. Plot showing squared 2-Wasserstein metric between power posterior and prior against the power posterior parameter $\gamma \in [0, \frac{1}{2}]$ under the three different prior assumptions described in Table 4.4.

and the results are presented in Fig. 4.4. The squared Wasserstein distance converges more rapidly compared to the Normal normal case. The Wasserstein distance converges when γ is below 0.2. Similar to the normal normal case, the uncertainty is greatest at lower values of γ, especially when the subsample size is smaller as shown in Fig. 4.5.

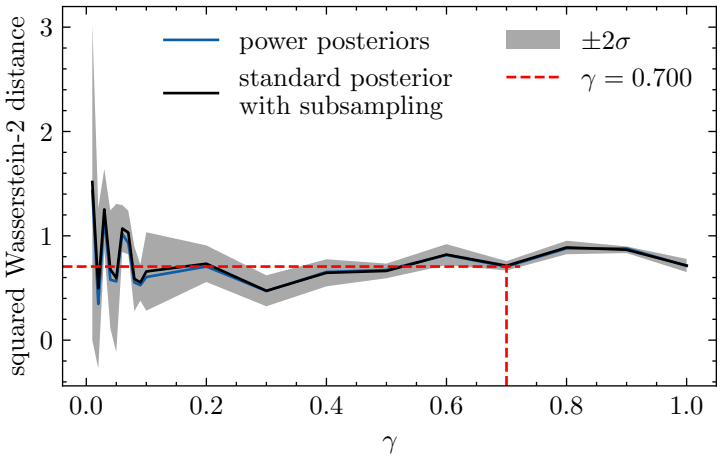


Figure 4.5: Comparison of the distances for power posteriors and standard posteriors with subsampling. The dashed lines represent the Wasserstein distance and the parameter γ at the saturated sample size at a 99% threshold. Poisson-Gamma conjugate case with unknown rate.

We also calculated the saturated sample size, which is presented in Table 4.5. The saturated sample size is reached at γ values of 0.50 for a 95% threshold and 0.70 for a 99% threshold.

Table 4.5: Saturated sample size and threshold for the Poisson Gamma model.

Wasserstein distance (W_t)	Threshold (CrI)	γ_{W_t}	Total sample size (n)	Saturated sample size (m_{sat})
0.672	95%	0.50	100	50
0.705	99%	0.70	100	70

This indicates that having 50 or 70 observations provides the same amount of information as the entire dataset.

4.4.4 Skew-normal distribution

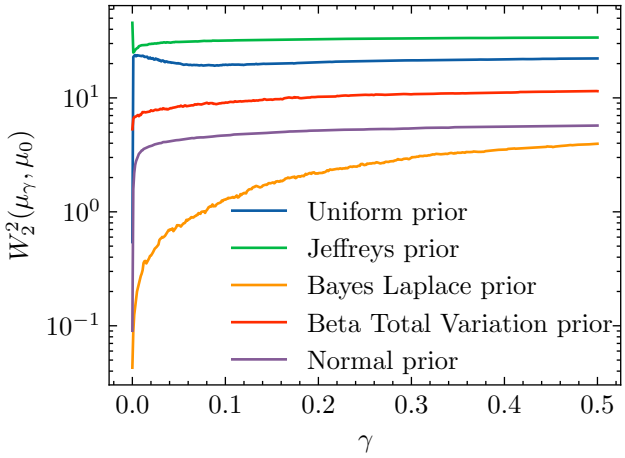
To gain more insight into prior-to-posterior transitions, we analysed the frontiers dataset found in the R package sn (Azzalini, 2023). Ghaderinezhad et al. (2022) previously analysed this dataset. The dataset is interesting because the maximum likelihood estimate of the coefficient of skewness lies on the boundary of the range [-0.995, 0.995] acceptable for the skew-normal family. The data was generated by drawing 50 samples from a skew-normal distribution with a skewness parameter 5.0.

$$f(x;\alpha) = \frac{2}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)\Phi\left(\alpha\frac{(x-\mu)}{\sigma}\right), \quad x \in \mathbb{R} \tag{4.11}$$

The density of a random variable x that follows a skew-normal distribution (Azzalini, 1985) is given by Eq. (4.11), where ϕ is the standard normal probability density function, Φ is the cumulative distribution function of the standard normal distribution, μ is the location, σ is the scale parameter and α is the skewness parameter. We fit the skew-normal distribution to the data with different sets of priors for the skewness parameter as in Ghaderinezhad et al. (2022). We compare the Wasserstein distance between power posteriors and the priors. The priors are Uniform prior, Jeffreys prior, Bayes-Laplace prior, Beta total variation prior (BTV), normal prior. The squared Wasserstein distance between the priors and the power posteriors are in Fig. 4.6. The distance between the prior ($\gamma = 0$) and the power posteriors increases with an increase in γ but becomes stable as γ approaches 1. However, unlike the conjugate cases, the distances do not converge to the same value as the plots are distinguishable.

The priors employed in this section are the same as those in (Ghaderinezhad et al., 2022). What differentiates this study is that we do not compute the WIM; instead, we compute the

Wasserstein distance between the priors and the power posteriors. The aim is to explore the evolution of the Wasserstein distance and how the priors transition to the posteriors for the skew-normal distribution.



‘

Figure 4.6: Squared Wasserstein distance between the prior and various power posteriors for different values $\gamma \in [0, 0.5]$ on the frontiers dataset.

Another way to visualise prior to posterior transitions is through density plots of the priors and the power posteriors. For the uniform prior Fig. 4.7, the mode is not visible, but as γ starts increasing, a mode becomes visible, and the uniform distribution starts transitioning into a skew distribution. On the other hand, the mode is visible for Jeffreys prior Fig. 4.8, but there is no skewness. As γ increases, the mode becomes more visible, and skewness becomes visible. As the value of γ increases, the mode becomes more peaked, and the range of the skewness parameter decreases. When $\gamma = 0$, the prior is less informative with no skewness. As γ increases, the prior develops a clear peak and becomes right-skewed. Skewness increases with increasing values of γ. This indicates the prior transitions from a less informative to an informative posterior (skew normal posterior) as γ increases. Like in the uniform flat prior, as γ increases, the prior peaks with less uncertainty and skewness to the right. The density plot of the prior ($\gamma = 0$) in Fig. 4.9 and the standard posterior ($\gamma = 1$) in Fig. 4.10 are shown for comparison with the power posteriors.

We now explain the importance of the concepts of power posteriors and saturated sample sizes. Power posteriors can be employed to specify prior distributions (O'Hagan, 1997). The power posteriors that occur before reaching the saturated sample size can be used to construct weakly informative priors. These priors allow the likelihood to dominate the posterior but still

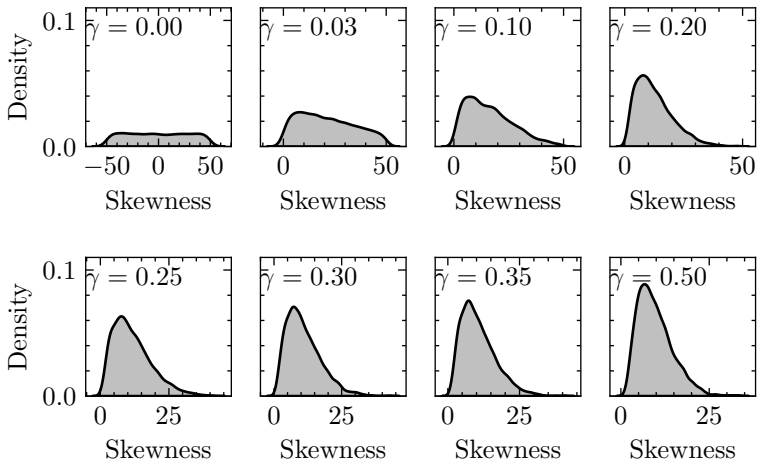


Figure 4.7: Marginal posteriors for the skewness parameter show the transition from a uniform prior ($\gamma = 0$) to various power posteriors ($0 < \gamma \leq \frac{1}{2}$). Posterior based on the frontier dataset for the skew-normal distribution.

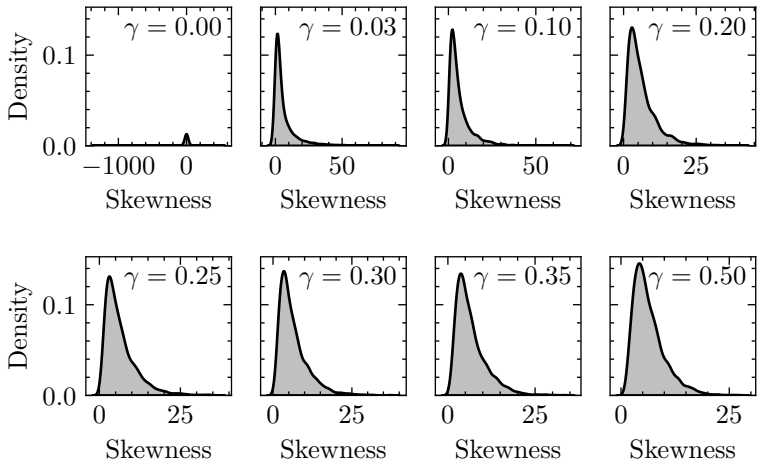


Figure 4.8: Marginal posteriors for the skewness parameter show the transition from Jeffreys prior ($\gamma = 0$) to various power posteriors ($0 < \gamma \leq \frac{1}{2}$). Posterior based on the frontier dataset for the skew-normal distribution.

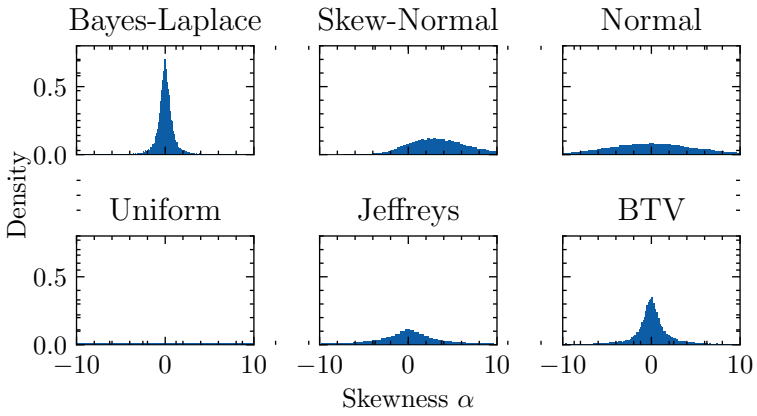


Figure 4.9: Frontier skew-normal: different priors. Beta total variation prior (BTV)

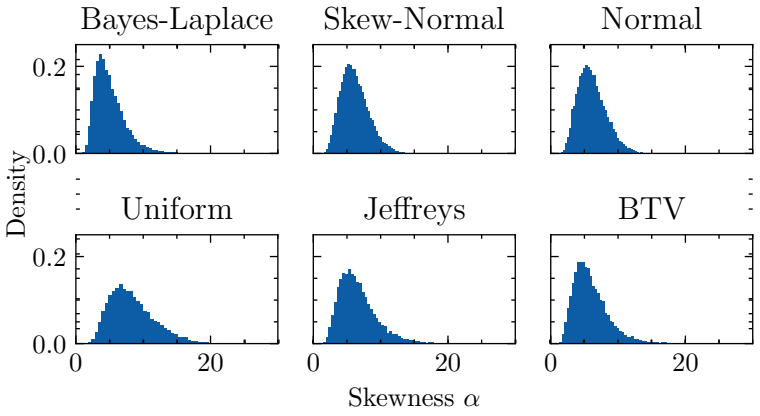


Figure 4.10: Frontier skew-normal: posteriors under different priors. This are standard posteriors for the different prior. The posteriors are all positively skewed unlike the priors in Fig. 4.9 but with different modes. Beta total variation prior (BTV).

provide minimal information about the parameters. This is because the Wasserstein distance between the prior and the power posteriors is highest at lower values of γ, implying that at lower values of γ, the prior has a high influence on the posterior.

Another important application of power posteriors is in determining appropriate sample sizes. For example, in clinical trials, interim analyses are conducted to adapt patient recruitment strategies based on the results of these analyses. One of these strategies is reducing the original sample size or recruiting more patients for a specific treatment arm (Berry, 2006; Ryan et al., 2022). The saturated sample size can be used to decide if recruiting more patients will make any difference or to stop patient recruitment. This is of enormous importance in the case of rare diseases where patient recruitment is difficult or if there are ethical concerns. The saturated sample size can also inform when to stop recruiting more survey participants.

4.5 Conclusion and discussions

In this study, we explored prior-to-posterior transitions using power posteriors. We have demonstrated that power posteriors weight the sample size, making them similar to subsampling techniques. The region from $\gamma = 0$ to $\gamma = 1$ forms a continuous path, transitioning smoothly from the prior distribution to the power posteriors, including the standard posterior. The distance between the prior and the standard posterior increases from $\gamma = 0$ to $\gamma = 1$. The power posterior converges to the standard posterior as the power value increases. For conjugate cases, the posterior distribution stabilises at lower power values, which we confirmed using the Wasserstein distance. Using density plots, we see that the prior transitions to power posteriors closer to the standard posterior at lower γ values suggesting that power posteriors can be a good alternative for parameter inference. They are currently used to handle model misspecification (Miller & Dunson, 2019).

We also provided derivations of power posteriors for conjugate cases. Additionally, power posteriors can be applied with lower power values to derive data-driven priors. This involves using a part of the data to create these priors and the remainder for the main analysis.

Furthermore, power posteriors allow for the use of improper priors with a subset of the data to derive proper priors, which are then utilised in subsequent analysis. This approach is similar to the fractional BF approach proposed by O'Hagan (1997), where updating an improper prior with a portion of the data helps refine it into a proper prior for the main analysis.

To conclude, we have also introduced the concept of saturated sample size, which has impli-

cations for clinical trials and survey statistics. The saturated sample size is the minimal number of observations with the same amount of information as the entire dataset. This saturated sample size occurs at a γ and power posterior. The power posterior corresponding to the value of γ at saturated sample size can be used for inference as it provides the same information as the standard posterior. It might be difficult to recruit more survey or clinical trial participants. In the case of clinical trials, an interim analysis can be done to determine if the saturated sample size has been reached. For survey statistics, the saturated sample size can be calculated as more participants complete the survey. The study can be concluded once the saturated sample size has been reached.

Code and data availability

The complete code used to generate the results will be available in a public repository upon submission.

Competing interests

The authors declare no competing interests.

Acknowledgements

This work was funded under the Luxembourg National Research Fund under the PRIDE programme (PRIDE17/12252781).

Appendix A

Proofs

A.1 Power posteriors for conjugate cases

Theorem 1. *Let m be a parameter normally distributed with prior mean m_0 and prior variance σ_0^2. The observed data is $x = (x_1, ..., x_n)$ with sample mean $\bar{x}$ and the data model is $x \overset{\text{iid}}{\sim} \mathcal{N}(m, \sigma^2)$. The power posterior can be written*

$$m \mid \bar{x}, \gamma \sim \mathcal{N}\left(\left(\frac{1}{\sigma_0^2} + \frac{\gamma n}{\sigma^2}\right)^{-1}\left(\frac{m_0}{\sigma_0^2} + \frac{\gamma n \bar{x}}{\sigma^2}\right), \left(\frac{1}{\sigma_0^2} + \frac{\gamma n}{\sigma^2}\right)^{-1}\right).$$

Proof. The standard posterior $\gamma = 1$ is a normal distribution and can be expressed as

$$m \mid \bar{x} \sim \mathcal{N}\left(\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}\left(\frac{m_0}{\sigma_0^2} + \frac{n\bar{x}}{\sigma^2}\right), \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}\right),$$

(see e.g. Gelman et al., 2020) for proof. Taking the normal likelihood and raising it to the power γ gives

$$\begin{aligned} p(\bar{x} \mid m)^\gamma &\propto \exp\left(-\frac{n}{2\sigma^2}(\bar{x} - m)^2\right)^\gamma \\ &\propto \exp\left(-\frac{\gamma n}{2\sigma^2}(\bar{x} - m)^2\right). \end{aligned}$$

Then, the power posterior can be derived by making the substitution $\sigma^2 \to \frac{\sigma^2}{\gamma}$ into the standard posterior. □

Theorem 2. *Let m be the mean and σ^2 be the unknown variance. The observed data is denoted as $x = (x_1, ..., x_n)$. The prior variance follows the inverse gamma distribution with shape pa-*

rameter α and scale parameter β. The data model is $x \overset{\text{iid}}{\sim} \mathcal{N}(m, \sigma^2)$. The power posterior can be written as

$$\sigma^2 \mid s^2, \gamma \sim \text{IG}\left(\alpha + \frac{\gamma n}{2}, \beta + \frac{\gamma n}{2} s^2\right),$$

where, $s^2 = (\sum_i (x_i - \mu)^2)/n$ is the sample variance.

Proof. The standard posterior $\gamma = 1$ is an inverse gamma and can be expressed as

$$\sigma^2 \mid s^2 \sim \text{IG}\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum (x_i - m)^2\right),$$

(see e.g. Gelman et al., 2020) for proof. Taking the normal likelihood and raising it to the power γ gives

$$p(x_1 \ldots, x_n | \sigma^2)^\gamma = (\sigma^2)^{-\frac{\gamma n}{2}} \exp\left(-\frac{\gamma}{2\sigma^2}\sum (x_i - m)^2\right).$$

This power posterior can be derived by inspection and substituting $\alpha + \frac{n}{2} \to \alpha + \frac{\gamma n}{2}$ and $\beta + \frac{1}{2}\sum (x_i - m)^2 \to \beta + \frac{\gamma}{2}\sum (x_i - m)^2$ in the standard posterior. □

Theorem 3. *Let λ be the rate parameter of the Poisson distributions, the observed data is $x = (x_1, \ldots, x_n)$ and the data model is $x \overset{\text{iid}}{\sim}$ Poisson(λ). When λ has a gamma prior with parameters α and β, the power posterior can be written as*

$$\lambda \mid \bar{x} \sim \text{Gamma}\left(\gamma n \bar{x} + \alpha, \gamma n + \beta\right).$$

Proof. The standard posterior $\gamma = 1$ is a gamma distribution and can be expressed as

$$\begin{aligned} \lambda \mid \bar{x} &\sim \text{Gamma}\left(n\bar{x} + \alpha, n + \beta\right) \\ \bar{x} &= \sum x_i / n, \end{aligned}$$

(see e.g. Gelman et al., 2020) for proof. Raise the likelihood to the power γ

$$p(x \mid \lambda)^\gamma = \frac{e^{-n\gamma\lambda}\lambda^{\sum x_i}}{\prod_{i=1}^n (x_i!)^\gamma} \tag{A.1}$$

The power posterior can be derived by substituting $\sum x_i + \alpha \to \gamma \sum x_i + \alpha$ and $n + \beta \to \gamma n + \beta$ in the standard posterior. □

Chapter 5

Conclusions and future work

5.1 Conclusions

Dynamical systems occur in various fields, such as epidemiology, engineering, ecology, and hydrology. These systems are often modelled using ordinary differential equations (ODEs), which vary in parameter counts. Models with more parameters tend to have higher variance and lower accuracy for out-of-sample data. Thus, a parsimonious model is preferred if it performs comparably to complex models for in-sample data. The Bayesian approach to model selection requires computing the marginal likelihood. However, calculating the marginal likelihood is computationally intensive because of the need for numerous likelihood evaluations. Therefore, information-theoretic criteria and marginal likelihood approximations are frequently used to bypass this computational challenge. These information-theoretic criteria may select different models based on specific criteria.

In conclusion, Chapter 2 presents advances in Bayesian model selection, focusing on dynamical systems, specifically those modelled by ODEs. By leveraging computational techniques, such as differentiable programming, replica exchange, and gradient-based Hamiltonian Monte Carlo with thermodynamic integration, the research offers a computationally efficient method for accurately computing marginal likelihoods. Besides, this thesis has shown that the marginal likelihood can differentiate between models in some cases when information-theoretic criteria do not provide a clear choice. The approach has been applied to synthetic and real data using ODE-based rainfall-runoff models. Source codes are publicly available to facilitate the adoption of these methods and enable their adoption in standard conceptual hydrological modelling toolboxes such as MARRMoT (Trotter et al., 2022) and SUPERFLEX (Dal Molin et al., 2021).

Moreover, the influence of the prior in Bayesian analysis on our conclusions or inference is well documented. However, while several formal prior impact assessment methods exist, none have been applied to ODEs. Typically, the priors for ODEs are strongly or weakly informative. Therefore, assessing the impact of priors on ODEs is essential since priors are informative. In addition, there is a lack of a clear interpretation of the metrics used by these methods. For instance, there is no threshold for saying that a prior has a higher or lower impact. Hence, in Chapter 3, the thesis extends the scope of prior impact assessments in Bayesian analysis to ODE models. Using computational optimal transport algorithms to calculate Wasserstein distances, the research introduces the prior scaled Wasserstein Impact Measure (sWIM), which provides a clearer relative interpretation of prior impacts. Thus, one can compare prior choices to see if they make any difference in the inference, in which case further justification of the chosen prior will be necessary. We have also exemplified the applicability of the sWIM for real-world data. In addition, our source code follows open science to facilitate the adoption of the technique.

Furthermore, the transition from prior to posterior is not well understood, underscoring the need for a better understanding of the influence of priors in Bayesian analysis. In Chapter 4, power posteriors and Wasserstein distances are used to gain insights into how the prior transitions to the posterior. Also, the concept of saturation sample size is introduced as an alternative to large sample sizes. This approach can shorten the completion time of studies without compromising the results. The thesis shows that power posteriors are equivalent to subsampling. There are also derivations for the power posteriors of conjugate cases.

5.1.1 Future work

The limitations of this thesis are on the application side rather than the methodology side. These limitations and recommendations for future work are listed below, highlighting the potential for further advancements.

1. The entire dataset for Magela Creek catchment in Australia was not used, which could potentially have improved the goodness of fit of the selected model. Only periods with complete data were used to avoid making assumptions about the nature of the missing values. Therefore, it is recommended to perform a Bayesian model selection using all available data. Additionally, model selection should be performed using complete case analysis by removing all missing values to evaluate the impact of missing data on model selection.

2. At present, it is not feasible to include domain-specific metrics such as KGE used in hydrology in the likelihood function (Cheng et al., 2014). Future research can explore approaches to directly include domain-specific metrics, such as KGE (Y. Liu et al., 2022) in hydrology, into the likelihood function and subsequent use in model selection.

3. The sWIM has clear interpretations below, at, and above one. However, the precise threshold for sWIM values remains to be determined. For instance, a sWIM of 0.50 and 0.90 have the same interpretation, indicating the need for more precise guidelines.

4. Power posteriors are used for robust inference and to set up power priors, but the optimal value for the power (e.g., $\gamma = 0.5$ or $\gamma = 0.1$) for posterior summary statistics remains undetermined, with the current standard being $\gamma = 1$. Also, the optimal value for power priors is yet to be determined.

References

Ali, S. M., & Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, *28*(1), 131–142. doi: 10.1111/j.2517-6161.1966.tb00626.x

Allanach, B. C., & Lester, C. G. (2008). Sampling using a 'bank' of clues. *Computer Physics Communications*, *179*(4), 256–266. doi: 10.1016/j.cpc.2008.02.020

Arrowsmith, D. K., & Place, C. M. (1990). *An introduction to dynamical systems*. Cambridge University Press.

Ashton, G., Bernstein, N., Buchner, J., Chen, X., Csányi, G., Fowlie, A., ... Yallup, D. (2022). Nested sampling for physical scientists. *Nature Reviews Methods Primers*, *2*(1), 39. doi: 10.1038/s43586-022-00121-x

Awrejcewicz, J. (2014). *Ordinary differential equations and mechanical systems*. Cham: Springer International Publishing. doi: 10.1007/978-3-319-07659-1

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, *12*(2), 171–178. Retrieved from `http://www.jstor.org/stable/4615982`

Azzalini, A. (2023). R package 'sn'. *The skew-normal and related distributions such as the skew-t and the SUN*. Retrieved from `http://azzalini.stat.unipd.it/SN/`

Bai, Z. D., Rao, C. R., & Wu, Y. (1999). Model selection with data-oriented penalty. *Journal of Statistical Planning and Inference*, *77*(1), 103–117. doi: 10.1016/S0378-3758(98)00168-2

Berger, J. O., Bernardo, J. M., & Sun, D. (2015). Overall objective priors. *Bayesian Analysis*, *10*(1). doi: 10.1214/14-BA915

Berger, J. O., & Pericchi, L. R. (1996). The intrinsic bayes factor for model selection and prediction. *Journal of the American Statistical Association*, *91*(433), 109-122. doi: 10.1080/01621459.1996.10476668

Berger, J. O., Pericchi, L. R., Ghosh, J. K., Samanta, T., & De Santis, F. (2001). Objective Bayesian methods for model selection: Introduction and comparison. *Lecture Notes-Monograph Series*, *38*, 135–207. Retrieved from `http://www.jstor.org/stable/4356165`

Bergström, S. (1976). *Development and application of a conceptual runoff model for Scandinavian catchments* (Tech. Rep. No. 7). Swedish Meteorological and Hydrological Institute.

Bergström, S., & Lindström, G. (2015). Interpretation of runoff processes in hydrological modelling—experience from the HBV approach. *Hydrological Processes*, *29*(16), 3535–3545. doi: 10.1002/hyp.10510

Berry, D. A. (2006). Bayesian clinical trials. *Nature Reviews Drug Discovery*, *5*(1), 27–36. doi: 10.1038/nrd1927

Beskos, A., Pillai, N., Roberts, G., Sanz-Serna, J.-M., & Stuart, A. (2013). Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli*, *19*, 1501–1532. doi: 10.3150/12-BEJ414

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*. doi: 10.48550/arXiv.1701.02434

BEVEN, K. (1997). Topmodel: A critique. *Hydrological Processes*, *11*(9), 1069-1085. doi: 10.1002/(SICI)1099-1085(199707)11:9<1069::AID-HYP545>3.0.CO;2-O

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., ... Goodman, N. D. (2019). Pyro: Deep universal probabilistic programming. *Journal*

of Machine Learning Research, *20*, 1–6. Retrieved from `http://jmlr.org/papers/v20/18-403.html`

Birgé, L., & Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields*, *138*(1), 33–73. doi: 10.1007/s00440-006-0011-8

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., ... Zhang, Q. (2018). *Jax: Composable transformations of python+numpy programs (version 0.3.13)*. Retrieved from `http://github.com/google/jax`

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455. doi: 10.1080/10618600.1998.10474787

Brunetti, C., Bianchi, M., Pirot, G., & Linde, N. (2019). Hydrogeological model selection among complex spatial priors. *Water Resources Research*, *55*(8), 6729–6753. doi: 10.1029/2019WR024840

Brunetti, C., & Linde, N. (2018). Impact of petrophysical uncertainty on Bayesian hydrogeophysical inversion and model selection. *Advances in Water Resources*, *111*, 346–359. doi: 10.1016/j.advwatres.2017.11.028

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

Burnham, K. P., & Anderson, D. R. (2002). Information and likelihood theory: A basis for model selection and inference. In K. P. Burnham & D. R. Anderson (Eds.), *Model selection and multimodel inference: A practical information-theoretic approach* (pp. 49–97). Springer. doi: 10.1007/978-0-387-22456-5_2

Calderhead, B., & Girolami, M. (2009). Estimating bayes factors via thermodynamic integration and population mcmc. *Computational Statistics & Data Analysis*, *53*(12), 4028–4045. doi: 10.1016/j.csda.2009.07.025

Calderhead, B., & Girolami, M. (2011). Statistical analysis of nonlinear dynamical systems using differential geometric sampling methods. *Interface Focus*, *1*(6), 821–835. doi: 10.1098/rsfs.2011.0051

Cao, T., Zeng, X., Wu, J., Wang, D., Sun, Y., Zhu, X., ... Long, Y. (2019). Groundwater contaminant source identification via Bayesian model selection and uncertainty quantification. *Hydrogeology Journal*, *27*(8), 2907–2918. doi: 10.1007/s10040-019-02055-3

Cappé, O., Guillin, A., Marin, J. M., & Robert, C. P. (2004). Population Monte Carlo. *Journal of Computational and Graphical Statistics*, *13*(4), 907–929. doi: 10.1198/106186004X12803

Carpenter, B. (2018). *Predator-prey population dynamics: The Lotka-Volterra model in Stan.* Retrieved 2023-07-13, from `https://mc-stan.org/users/documentation/case-studies/lotka-volterra-predator-prey.html` (Stan case studies)

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). *Stan*: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. doi: 10.18637/jss.v076.i01

Casella, G., Girón, F. J., Martínez, M. L., & Moreno, E. (2009). Consistency of Bayesian procedures for variable selection. *The Annals of Statistics*, *37*(3), 1207–1228. doi: 10.1214/08-AOS606

Chen, M.-H., Shao, Q.-M., & Ibrahim, J. G. (2000). Estimating ratios of normalizing constants. In M.-H. Chen, Q.-M. Shao, & J. G. Ibrahim (Eds.), *Monte Carlo methods in Bayesian computation* (pp. 124–190). Springer New York. doi: 10.1007/978-1-4612-1276-8_5

Chen, R. T., Rubanova, Y., Bettencourt, J., & Duvenaud, D. K. (2018). Neural ordinary differential equations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 31* (pp. 6571–6583). Curran Associates, Inc.

Cheng, Q.-B., Chen, X., Xu, C.-Y., Reinhardt-Imjela, C., & Schulte, A. (2014, November). Improvement and comparison of likelihood functions for model calibration and

parameter uncertainty analysis within a Markov chain Monte Carlo scheme. *Journal of Hydrology*, *519*, 2202–2214. doi: 10.1016/j.jhydrol.2014.10.008

Chib, S. (1995). Marginal likelihood from the gibbs output. *Journal of the American Statistical Association*, *90*(432), 1313–1321. doi: 10.1080/01621459.1995.10476635

Chib, S., & Jeliazkov, I. (2001). Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association*, *96*(453), 270–281. doi: 10 .1198/016214501750332848

Chib, S., & Kuffner, T. A. (2016). Bayes factor consistency. *arXiv preprint arXiv:1607.00292*. doi: 10.48550/ARXIV.1607.00292

Chopin, N., & Robert, C. P. (2010). Properties of nested sampling. *Biometrika*, *97*(3), 741–755. doi: 10.1093/biomet/asq021

Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, *91*(434), 883–904. doi: 10.1080/01621459.1996.10476956

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 26). Curran Associates, Inc.

Cuturi, M., & Doucet, A. (2014). Fast computation of wasserstein barycenters. In E. P. Xing & T. Jebara (Eds.), *Proceedings of the 31st International Conference on Machine Learning* (Vol. 32, pp. 685–693). Bejing, China: PMLR. Retrieved from `https://proceedings.mlr.press/v32/cuturi14.html`

Cuturi, M., Klein, M., & Ablin, P. (2023). Monge, Bregman and Occam: Interpretable optimal transport in high-dimensions with feature-sparse maps. *preprint arXiv:2302.04065*. doi: 10.48550/ARXIV.2302.04065

Cuturi, M., Meng-Papaxanthos, L., Tian, Y., Bunne, C., Davis, G., & Teboul, O. (2022). Optimal Transport Tools (OTT): A Jax toolbox for all things Wasserstein. *arXiv preprint arXiv:2201.12324*.

Dai, C., & Liu, J. S. (2022). Monte Carlo approximation of Bayes factors via mixing with surrogate distributions. *Journal of the American Statistical Association*, *117*(538), 765–780. doi: 10.1080/01621459.2020.1811100

Dal Molin, M., Kavetski, D., & Fenicia, F. (2021). Superflexpy 1.3.0: An open-source python framework for building, testing, and improving conceptual hydrological models. *Geoscientific Model Development*, *14*(11), 7047–7072. doi: 10.5194/gmd-14-7047-2021

Davarci, O. O., Yang, E. Y., Viguerie, A., Yankeelov, T. E., & Lorenzo, G. (2024). Dynamic parameterization of a modified SEIRD model to analyze and forecast the dynamics of COVID-19 outbreaks in the United States. *Engineering with Computers*, *40*(2), 813–837. doi: 10.1007/s00366-023-01816-9

DeepMind, Babuschkin, I., Baumli, K., Bell, A., Bhupatiraju, S., Bruce, J., ... Viola, F. (2020). *The DeepMind JAX Ecosystem.* Retrieved from `http://github.com/deepmind`

Devia, G. K., Ganasri, B., & Dwarakish, G. (2015). A review on hydrological models. *Aquatic Procedia*, *4*, 1001–1007. (International Conference on Water Resources, Coastal and Ocean Engineering (ICWRCOE'15)) doi: 10.1016/j.aqpro.2015.02.126

Dillon, J. V., Langmore, I., Tran, D., Brevdo, E., Vasudevan, S., Moore, D., ... Saurous, R. A. (2017). Tensorflow distributions. *preprint arXiv:1711.10604*. doi: 10.48550/ARXIV.1711.10604

Doroudi, S. (2020). The bias-variance tradeoff: How data science can inform educational debates. *AERA Open*, *6*(4). doi: 10.1177/2332858420977208

Drineas, P., Mahoney, M. W., & Muthukrishnan, S. (2006). Sampling algorithms for ℓ_2 regression and applications. In *Proceedings of the seventeenth annual ACM-SIAM symposium on discrete algorithms* (pp. 1127–1136). Miami, Florida: Society for Industrial and Applied Mathematics.

Duan, Q., Sorooshian, S., & Gupta, V. (1992). Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resources Research*, *28*(4), 1015-1031. doi: 10.1029/91WR02985

Duane, S., Kennedy, A., Pendleton, B. J., & Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, *195*(2), 216–222. doi: 10.1016/0370-2693(87)91197-X

Earl, D. J., & Deem, M. W. (2005). Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, *7*(23), 3910. doi: 10.1039/b509983h

Elsheikh, A. H., Wheeler, M. F., & Hoteit, I. (2014). Hybrid nested sampling algorithm for Bayesian model selection applied to inverse subsurface flow problems. *Journal of Computational Physics*, *258*, 319–337. doi: 10.1016/j.jcp.2013.10.001

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

Feroz, F., Hobson, M. P., & Bridges, M. (2009). MultiNest: An efficient and robust Bayesian inference tool for cosmology and particle physics. *Monthly Notices of the Royal Astronomical Society*, *398*(4), 1601–1614. doi: 10.1111/j.1365-2966.2009.14548.x

Fouskakis, D., Ntzoufras, I., & Draper, D. (2015). Power-expected-posterior priors for variable selection in Gaussian linear models. *Bayesian Analysis*, *10*(1), 75 – 107. doi: 10.1214/14-BA887

Frasso, G., & Lambert, P. (2016). Bayesian inference in an extended SEIR model with nonparametric disease transmission rate: An application to the Ebola epidemic in Sierra Leone. *Biostatistics (Oxford, England)*, *17*(4), 779–792. doi: 10.1093/biostatistics/kxw027

Friel, N., & Pettitt, A. N. (2008). Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(3), 589–607. doi: 10.1111/j.1467-9868.2007.00650.x

Frostig, R., Johnson, M. J., & Leary, C. (2018). Compiling machine learning programs via high-level tracing. *Systems for Machine Learning*, *4*(9).

Gelfand, A. E., & Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B (Methodological)*, *56*(3), 501–514. doi: 10.1111/j.2517-6161.1994.tb01996.x

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*(3), 515–534. doi: 10.1214/06-BA117A

Gelman, A. (2013). Two simple examples for understanding posterior p-values whose distributions are far from uniform. *Electronic Journal of Statistics*, *7*.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (Third ed.). CRC Press, Taylor and Francis Group. doi: 10.1201/b16018

Gelman, A., & Meng, X.-L. (1998). Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statistical Science*, *13*(2). doi: 10.1214/ss/1028905934

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, 733–760.

Gelman, A., Roberts, G. O., & Gilks, W. R. (1996). Efficient Metropolis jumping rules. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics* (p. 599-608). Oxford University Press, Oxford.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. doi: 10.1214/ss/1177011136

Gelman, A., Simpson, D., & Betancourt, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy*, *19*(10). doi: 10.3390/e19100555

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., ... Modrák, M. (2020). Bayesian workflow. *arXiv*. doi: 10.48550/ARXIV.2011.01808

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*(1), 1–58. doi: 10.1162/neco.1992.4.1.1

Geweke, J. (1991). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments* (Tech. Rep.). Federal Reserve Bank of Minneapolis.

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. *Interface Foundation of North America*. Retrieved from `https://hdl.handle.net/11299/58440`

Geyer, C. J. (1992). Practical Markov Chain Monte Carlo. *Statistical Science*, *7*(4), 473–483.

Ghaderinezhad, F., Ley, C., & Serrien, B. (2022). The Wasserstein Impact Measure (WIM): A practical tool for quantifying prior impact in Bayesian statistics. *Computational Statistics & Data Analysis*, *174*, 107352. doi: 10.1016/j.csda.2021.107352

Ghorbanzadeh, D., Jaupi, L., & Durand, P. (2014). A method to simulate the skew normal distribution. *Applied Mathematics*, *05*(13), 2073–2076. doi: 10.4236/am.2014.513201

Ghosal, P., Nutz, M., & Bernton, E. (2022). Stability of entropic optimal transport and Schrödinger bridges. *Journal of Functional Analysis*, *283*(9), 109622. doi: 10.1016/j.jfa.2022.109622

Ghosh, M. (2011). Objective priors: An introduction for frequentists. *Statistical Science*, *26*(2). doi: 10.1214/10-STS338

Gibson, G. C., Reich, N. G., & Sheldon, D. (2023). Real-time mechanistic Bayesian forecasts of COVID-19 mortality. *The Annals of Applied Statistics*, *17*(3). doi: 10.1214/22-AOAS1671

Girolami, M. (2008, November). Bayesian inference for differential equations. *Theoretical Computer Science*, *408*(1), 4–16. doi: 10.1016/j.tcs.2008.07.005

Girolami, M., & Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *73*(2), 123–214. doi: 10.1111/j.1467-9868.2010.00765.x

Gneiting, T., & Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. doi: 10.1198/016214506000001437

Golchi, S. (2019). Informative priors in Bayesian inference and computation. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *12*(2), 45–55. doi: 10 .1002/sam.11371

Grinsztajn, L., Semenova, E., Margossian, C. C., & Riou, J. (2021). Bayesian workflow for disease transmission modeling in Stan. *Statistics in Medicine*, *40*(27), 6209–6234. doi: 10.1002/sim.9164

Gustafson, P. (1996). Local sensitivity of inferences to prior marginals. *Journal of the American Statistical Association*, *91*(434), 774–781. doi: 10.1080/01621459.1996 .10476945

Hanbing Xu, J. L., Songbai Song, & Guo, T. (2023). Hybrid model for daily runoff interval predictions based on Bayesian inference. *Hydrological Sciences Journal*, *68*(1), 62–75. doi: 10.1080/02626667.2022.2145201

Hansmann, U. H. (1997). Parallel tempering algorithm for conformational studies of biological molecules. *Chemical Physics Letters*, *281*(1-3), 140–150. doi: 10.1016/ S0009-2614(97)01198-6

He, S., Peng, Y., & Sun, K. (2020). SEIR modeling of the COVID-19 and its dynamics. *Nonlinear Dynamics*, *101*(3), 1667–1680. doi: 10.1007/s11071-020-05743-y

Henderson, R. W., & Goggans, P. M. (2019). TI-Stan: Model comparison using thermodynamic integration and HMC. *Entropy*, *21*(12), 1161. doi: 10.3390/e21121161

Hjort, N. L., Dahl, F. A., & Steinbakk, G. H. (2006). Post-processing posterior predictive p values. *Journal of the American Statistical Association*, *101*(475), 1157–1174. doi: 10.1198/016214505000001393

Hoffman, M. D., & Gelman. (2014). The No-U-Turn Sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

Hug, S., Schmidl, D., Li, W. B., Greiter, M. B., & Theis, F. J. (2016). Bayesian model selection methods and their application to biological ODE systems. In *Uncertainty in biology* (pp. 243–268). Springer.

Hung, H. M. J., O'Neill, R. T., Bauer, P., & Kohne, K. (1997). The behavior of the p-value when the alternative hypothesis is true. *Biometrics*, *53*(1), 11. doi: 10.2307/2533093

Höge, M., Guthke, A., & Nowak, W. (2019, May). The hydrologist's guide to Bayesian model selection, averaging and combination. *Journal of Hydrology*, *572*, 96–107. doi: 10.1016/j.jhydrol.2019.01.072

Höge, M., Scheidegger, A., Baity-Jesi, M., Albert, C., & Fenicia, F. (2022). Improving hydrologic models for predictions and process understanding using neural ODEs. *Hydrology and Earth System Sciences*, *26*(19), 5085–5102. doi: 10.5194/hess-26-5085-2022

Höge, M., Wöhling, T., & Nowak, W. (2018). A primer for model selection: the decisive role of model complexity. *Water Resources Research*, *54*(3), 1688-1715. doi: 10.1002/2017WR021902

Iba, Y. (2000). Population Monte Carlo algorithms. *Transactions of the Japanese Society for Artificial Intelligence*, *16*(2), 279–286.

Ibrahim, J. G., & Chen, M.-H. (2000). Power prior distributions for regression models. *Statistical Science*, *15*(1), 46–60. Retrieved 2024-07-22, from `http://www.jstor.org/stable/2676676`

Information and likelihood theory: A basis for model selection and inference. (2002). In K. P. Burnham & D. R. Anderson (Eds.), *Model selection and multimodel inference: A practical information-theoretic approach* (pp. 49–97). New York, NY: Springer New York. doi: 10.1007/978-0-387-22456-5_2

Jansen, K. F., Teuling, A. J., Craig, J. R., Dal Molin, M., Knoben, W. J. M., Parajka, J., ... Melsen, L. A. (2021). Mimicry of a conceptual hydrological model (HBV): What's in a name? *Water Resources Research*, *57*(5). doi: 10.1029/2020WR029143

Jones, D. E., Trangucci, R. N., & Chen, Y. (2022). Quantifying observed prior impact. *Bayesian Analysis*, *17*(3), 737–764. doi: 10.1214/21-BA1271

Kamrujjaman, M., Saha, P., Islam, M. S., & Ghosh, U. (2022). Dynamics of SEIR model: A case study of COVID-19 in Italy. *Results in Control and Optimization*, *7*, 100119. doi: 10.1016/j.rico.2022.100119

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

Kavetski, D., & Clark, M. P. (2011). Numerical troubles in conceptual hydrology: Approximations, absurdities and impact on hypothesis testing. *Hydrological Processes*, *25*(4), 661–670. doi: 10.1002/hyp.7899

Kavetski, D., & Fenicia, F. (2011). Elements of a flexible approach for conceptual hydrological modeling: 2. application and experimental insights. *Water Resources Research*, *47*(11). doi: 10.1029/2011WR010748

Kelly, J., Bettencourt, J., Johnson, M. J., & Duvenaud, D. K. (2020). Learning differential equations that are easy to solve. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 4370–4380). Curran Associates, Inc. Retrieved from `https://proceedings.neurips.cc/paper_files/paper/2020/file/2e255d2d6bf9bb33030246d31f1a79ca-Paper.pdf`

Kemp, F., Proverbio, D., Aalto, A., Mombaerts, L., Fouquier d'Hérouël, A., Husch, A., ... Magni, S. (2021). Modelling COVID-19 dynamics and potential for herd immunity by vaccination in Austria, Luxembourg and Sweden. *Journal of Theoretical Biology*, *530*, 110874. doi: 10.1016/j.jtbi.2021.110874

Kendall, W. S., Liang, F., & Wang, J.-S. (2005). *Markov chain Monte Carlo: innovations and applications* (Vol. 7). World Scientific.

Kennedy, M. C., & O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(3), 425-464. doi: 10.1111/1467-9868.00294

Kidger, P. (2021). *On Neural Differential Equations* (Unpublished doctoral dissertation). University of Oxford.

Kidger, P., Morrill, J., Foster, J., & Lyons, T. (2020). Neural controlled differential equations for irregular time series. *Advances in Neural Information Processing Systems*, *33*, 6696–6707. Retrieved from `https://proceedings.neurips.cc/paper_files/paper/2020/file/4a5876b450b45371f6cfe5047ac8cd45-Paper.pdf`

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

Krapu, C., & Borsuk, M. (2022). A differentiable hydrology approach for modeling with time-varying parameters. *Water Resources Research*, *58*(9). doi: 10.1029/2021WR031377

Lai, C.-C., Hsu, C.-Y., Jen, H.-H., Yen, A. M.-F., Chan, C.-C., & Chen, H.-H. (2021). The Bayesian Susceptible-Exposed-Infected-Recovered model for the outbreak of COVID-19 on the Diamond Princess Cruise Ship. *Stochastic Environmental Research and Risk Assessment*, *35*(7), 1319–1333. doi: 10.1007/s00477-020-01968-w

Laloy, E., & Vrugt, J. A. (2012). High-dimensional posterior exploration of hydrologic models using multiple-try DREAM $_{(ZS)}$ and high-performance computing: Efficient MCMC for high-dimensional problems. *Water Resour. Res.*, *48*(1). doi: 10.1029/2011WR010608

Lartillot, N., & Philippe, H. (2006). Computing Bayes factors using thermodynamic integration. *Systematic Biology*, *55*(2), 195–207. doi: 10.1080/10635150500433722

Lenk, P. J., & DeSarbo, W. S. (2000). Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika*, *65*(1), 93–119. doi: 10.1007/BF02294188

Li, M. Y., Graef, J. R., Wang, L., & Karsai, J. (1999). Global dynamics of a SEIR model with varying total population size. *Mathematical Biosciences*, *160*(2), 191–213. doi: 10.1016/S0025-5564(99)00030-9

Li, M. Y., & Muldowney, J. S. (1995). Global stability for the SEIR model in epidemiology. *Mathematical Biosciences*, *125*(2), 155–164. doi: 10.1016/0025-5564(95)92756-5

Liu, P., Elshall, A. S., Ye, M., Beerli, P., Zeng, X., Lu, D., & Tao, Y. (2016). Evaluating marginal likelihood with thermodynamic integration method and comparison with several other numerical methods. *Water Resources Research*, *52*(2), 734–758. doi: 10.1002/2014WR016718

Liu, S., She, D., Zhang, L., & Xia, J. (2023). An improved approximate Bayesian computation approach for high-dimensional posterior exploration of hydrological models. *Hydrology and Earth System Sciences Discussions*, 1–46. (Publisher: Copernicus GmbH) doi: 10.5194/hess-2022-414

Liu, Y., Fernández-Ortega, J., Mudarra, M., & Hartmann, A. (2022). Pitfalls and a feasible solution for using KGE as an informal likelihood function in MCMC methods: $\text{DREAM}_{(\text{zs})}$ as an example. *Hydrology and Earth System Sciences*, *26*(20), 5341–5355. (Publisher: Copernicus GmbH) doi: 10.5194/hess-26-5341-2022

Llorente, F., Martino, L., Delgado, D., & López-Santiago, J. (2023). Marginal likelihood computation for model selection and hypothesis testing: an extensive review. *SIAM Review*, *65*(1), 3–58. doi: 10.1137/20M1310849

Ma, P., Mahoney, M., & Yu, B. (2014). A statistical perspective on algorithmic leveraging. In E. P. Xing & T. Jebara (Eds.), *Proceedings of the 31st International Conference on Machine Learning* (Vol. 32, pp. 91–99). Bejing, China: PMLR.

Machac, D., Reichert, P., & Albert, C. (2016). Emulation of dynamic simulators with application to hydrology. *Journal of Computational Physics*, *313*, 352-366. doi: 10 .1016/j.jcp.2016.02.046

MacKay, D. J. (1992). Bayesian interpolation. *Neural Computation*, *4*(3), 415–447. doi: 10.1162/neco.1992.4.3.415

MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

Margossian, C. C. (2019). A review of automatic differentiation and its efficient implementation. *WIREs Data Mining and Knowledge Discovery*, *9*(4). doi: 10.1002/widm.1305

Marshall, L., Nott, D., & Sharma, A. (2005). Hydrological model selection: A Bayesian alternative. *Water Resources Research*, *41*(10). doi: 10.1029/2004WR003719

Mathieu, E., Ritchie, H., Rodés-Guirao, L., Appel, C., Giattino, C., Hasell, J., ... Roser, M. (2020). *Coronavirus pandemic (COVID-19).* Retrieved 01-06-2023, from `https://covid.ourworldindata.org/data/owid-covid-data.csv`

Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics*, *22*(3). doi: 10.1214/aos/1176325622

Meng, X.-L., & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, 831–860.

Merz, R., Parajka, J., & Blöschl, G. (2009). Scale effects in conceptual hydrological modeling. *Water Resources Research*, *45*(9). doi: 10.1029/2009WR007872

Miasojedow, B., Moulines, E., & Vihola, M. (2013). An adaptive parallel tempering algorithm. *Journal of Computational and Graphical Statistics*, *22*(3), 649–664. doi: 10.1080/10618600.2013.778779

Miles, P. (2019). Pymcmcstat: a python package for Bayesian inference using delayed rejection adaptive Metropolis. *Journal of Open Source Software*, *4*(38), 1417. doi: 10.21105/joss.01417

Miller, J. W., & Dunson, D. B. (2019). Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, *114*(527), 1113–1125. (PMID: 31942084) doi: 10.1080/01621459.2018.1469995

Mingas, G., & Bouganis, C.-S. (2016). Population-based MCMC on multi-core CPUs, GPUs and FPGAs. *IEEE Transactions on Computers*, *65*(4), 1283–1296. doi: 10.1109/TC.2015.2439256

Mingo, D. N., & Hale, J. S. (2024). *Wasserstein distance prior impact assessment for ODE models.* Zenodo. doi: 10.5281/zenodo.11553775

Mingo, N. D., & Hale, J. S. (2023, November). *Selecting a conceptual hydrological model using bayes' factors (version v1.0.0).* Zenodo. doi: 10.5281/zenodo.10202093

Moore, R. E., Rosato, C., & Maskell, S. (2022). Refining epidemiological forecasts with simple scoring rules. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *380*(2233), 20210305. doi: 10.1098/rsta.2021.0305

Morita, S., Thall, P. F., & Müller, P. (2008). Determining the effective sample size of a parametric prior. *Biometrics*, *64*(2), 595–602. doi: 10.1111/j.1541-0420.2007.00888.x

Mwalili, S., Kimathi, M., Ojiambo, V., Gathungu, D., & Mbogo, R. (2020). SEIR model for COVID-19 dynamics incorporating the environment and social distancing. *BMC Research Notes*, *13*(1), 352. doi: 10.1186/s13104-020-05192-1

Newton, M. A., & Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B (Methodological)*, *56*(1), 3–26. doi: 10.1111/j.2517-6161.1994.tb01956.x

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

Nur, D., Allingham, D., Rousseau, J., Mengersen, K. L., & McVinish, R. (2009). Bayesian hidden Markov model for DNA sequence segmentation: A prior sensitivity analysis. *Statistical Genetics & Statistical Genomics: Where Biology, Epistemology, Statistics, and Computation Collide*, *53*(5), 1873–1882. doi: 10.1016/j.csda.2008.07.007

Ogata, Y. (1989). A Monte Carlo method for high dimensional integration. *Numerische Mathematik*, *55*(2), 137–157. doi: 10.1007/BF01406511

Ogata, Y. (1990). A Monte Carlo method for an objective Bayesian procedure. *Annals of the Institute of Statistical Mathematics*, *42*(3), 403–433. doi: 10.1007/BF00049299

O'Hagan, A. (1997). Properties of intrinsic and fractional Bayes factors. *Test*, *6*(1), 101–118. doi: 10.1007/BF02564428

Panaretos, V. M., & Zemel, Y. (2019). Statistical aspects of Wasserstein distances. *Annual review of statistics and its application*, *6*, 405–431.

Parajka, J., Merz, R., & Blöschl, G. (2007). Uncertainty and multiple objective calibration in regional water balance modelling: case study in 320 Austrian catchments. *Hydrological Processes*, *21*(4), 435–446. doi: 10.1002/hyp.6253

Pedroza, C., Han, W., Thanh Truong, V. T., Green, C., & Tyson, J. E. (2018). Performance of informative priors skeptical of large treatment effects in clinical trials: a simulation study. *Statistical Methods in Medical Research*, *27*(1), 79–96. doi: 10.1177/0962280215620828

Peyré, G., & Cuturi, M. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, *11*(5-6), 355–607.

Quintero, Y., Ardila, D., Camargo, E., Rivas, F., & Aguilar, J. (2021). Machine learning models for the prediction of the SEIRD variables for the COVID-19 pandemic based on a deep dependence analysis of variables. *Computers in Biology and Medicine*, *134*, 104500. doi: 10.1016/j.compbiomed.2021.104500

Rackauckas, C., Ma, Y., Martensen, J., Warner, C., Zubov, K., Supekar, R., ... Edelman, A. (2020). Universal differential equations for scientific machine learning. *arXiv*. doi: 10.48550/ARXIV.2001.04385

Radford M. Neal. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo.* CRC Press. doi: 10.1201/b10905-7

Roberts, G. O., & Rosenthal, J. S. (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, *18*(2), 349–367. doi: 10.1198/jcgs.2009.06134

Roos, M., & Held, L. (2011). Sensitivity analysis in Bayesian generalized linear mixed models for binary data. *Bayesian Analysis*, *6*(2), 259 – 278. doi: 10.1214/11-BA609

Roos, M., Martins, T. G., Held, L., & Rue, H. (2015). Sensitivity analysis for Bayesian hierarchical models. *Bayesian Analysis*, *10*(2). doi: 10.1214/14-BA909

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 1151–1172.

Ryan, E. G., Couturier, D.-L., & Heritier, S. (2022). Bayesian adaptive clinical trial designs for respiratory medicine. *Respirology*, *27*(10), 834-843. doi: 10.1111/resp .14337

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016, apr). Probabilistic programming in python using PyMC3. *PeerJ Computer Science*, *2*, e55. doi: 10.7717/peerj-cs.55

Scetbon, M., Cuturi, M., & Peyré, G. (2021). Low-rank sinkhorn factorization. In *International Conference on Machine Learning* (pp. 9344–9354). PMLR.

Schaffer, W. M. (1985). Order and chaos in ecological systems. *Ecology*, *66*(1), 93–106. doi: 10.2307/1941309

Schmidli, H., Gsteiger, S., Roychoudhury, S., O'Hagan, A., Spiegelhalter, D., & Neuenschwander, B. (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, *70*(4), 1023–1032. doi: 10.1111/biom.12242

Seibert, J., & Bergström, S. (2022). A retrospective on hydrological catchment modelling based on half a century with the HBV model. *Hydrology and Earth System Sciences*, *26*(5), 1371–1388. doi: 10.5194/hess-26-1371-2022

Shafii, M., Tolson, B., & Matott, L. S. (2014). Uncertainty-based multi-criteria calibration of rainfall-runoff models: A comparative study. *Stochastic Environmental Research and Risk Assessment*, *28*(6), 1493–1510. doi: 10.1007/s00477-014-0855-x

Shah, N. H., & Gupta, J. (2013). SEIR model and simulation for vector-borne diseases. *Applied Mathematics*, *4*(8), 13–17. doi: 10.4236/am.2013.48A003

Skilling, J. (2004). Nested Sampling. In *AIP Conference Proceedings* (Vol. 735, pp. 395–405). AIP. doi: 10.1063/1.1835238

Skilling, J. (2006). Nested sampling for general Bayesian computation. *Bayesian Analysis*, *1*(4). doi: 10.1214/06-BA127

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639. doi: 10.1111/1467-9868.00353

Stefan, A. M., Katsimpokis, D., Gronau, Q. F., & Wagenmakers, E.-J. (2022). Expert agreement in prior elicitation and its effects on Bayesian inference. *Psychonomic Bulletin & Review*, *29*(5), 1776–1794. doi: 10.3758/s13423-022-02074-4

Swendsen, R. H., & Wang, J.-S. (1986). Replica Monte Carlo simulation of spin-glasses. *Physical Review Letters*, *57*(21), 2607–2609. doi: 10.1103/PhysRevLett.57.2607

Tang, Y., Marshall, L., Sharma, A., & Smith, T. (2016). Tools for investigating the prior distribution in Bayesian hydrology. *Journal of Hydrology*, *538*, 551–562. doi: 10.1016/j.jhydrol.2016.04.032

Thijssen, B., Dijkstra, T. M. H., Heskes, T., & Wessels, L. F. A. (2016). BCM: Toolkit for Bayesian analysis of computational models using samplers. *BMC Systems Biology*, *10*(1), 100. doi: 10.1186/s12918-016-0339-3

Torrie, G., & Valleau, J. (1977). Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, *23*(2), 187–199. doi: 10.1016/0021-9991(77)90121-8

Trotter, L., Knoben, W. J. M., Fowler, K. J. A., Saft, M., & Peel, M. C. (2022). Modular Assessment of Rainfall–Runoff Models Toolbox (MARRMoT) v2.1: An object-oriented implementation of 47 established hydrological models for improved speed and readability. *Geoscientific Model Development*, *15*(16), 6359–6369. doi: 10.5194/gmd-15-6359-2022

Vallender, S. S. (1974). Calculation of the Wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, *18*(4), 784–786. (Publisher: Society for Industrial and Applied Mathematics) doi: 10.1137/1118101

Varrette, S., Cartiaux, H., Peter, S., Kieffer, E., Valette, T., & Olloh, A. (2022, July). Management of an academic HPC & research computing facility: the ULHPC expe-

rience 2.0. In *Proc. of the 6th ACM High Performance Computing and Cluster Technologies Conf. (HPCCT 2022).* Fuzhou, China: Association for Computing Machinery (ACM).

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. doi: 10.1007/s11222-016-9696-4

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: an improved $\hat{R}$ for assessing convergence of MCMC (with discussion). *Bayesian Analysis*, *16*(2). doi: 10.1214/20-BA1221

Villani, C. (2009). *Optimal transport: Old and new* (No. 338). Springer. doi: 10.1007/978-3-540-71050-9

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... SciPy 1.0 Contributors (2020). SciPy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, *17*, 261–272. doi: 10.1038/s41592-019-0686-2

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

Vousden, W. D., Farr, W. M., & Mandel, I. (2016). Dynamic temperature selection for parallel tempering in Markov chain Monte Carlo simulations. *Monthly Notices of the Royal Astronomical Society*, *455*(2), 1919–1937. doi: 10.1093/mnras/stv2422

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

Vrugt, J. A., Ter Braak, C. J. F., Clark, M. P., Hyman, J. M., & Robinson, B. A. (2008). Treatment of input uncertainty in hydrologic modeling: doing hydrology backward with Markov chain Monte Carlo simulation. *Water Resources Research*, *44*(12). doi: 10.1029/2007WR006720

Wang, Z., & Xu, X. (2021). Calibration of posterior predictive p-values for model checking. *Journal of Statistical Computation and Simulation*, *91*(6), 1212–1242. doi: 10.1080/00949655.2020.1844701

Watanabe, S., & Opper, M. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*(12).

Weiss, R. (1996). An approach to Bayesian sensitivity analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(4), 739–750. doi: 10.1111/j.2517-6161.1996.tb02112.x

Wiesenfarth, M., & Calderazzo, S. (2020). Quantification of prior impact in terms of effective current sample size. *Biometrics*, *76*(1), 326–336. doi: 10.1111/biom.13124

Wood, E. F., Lettenmaier, D. P., & Zartarian, V. G. (1992). A land-surface hydrology parameterization with subgrid variability for general circulation models. *Journal of Geophysical Research: Atmospheres*, *97*(D3), 2717-2728. doi: 10.1029/91JD01786

Xifara, T., Sherlock, C., Livingstone, S., Byrne, S., & Girolami, M. (2014). Langevin diffusions and the Metropolis-adjusted Langevin algorithm. *Statistics & Probability Letters*, *91*, 14-19. doi: doi.org/10.1016/j.spl.2014.04.002

Xinyu, Z., Cao, J., & Carroll, R. J. (2015). On the selection of ordinary differential equation models with application to predator-prey dynamical models. *Biometrics*, *71*(1), 131–138. doi: 10.1111/biom.12243

Yao, Y., & Wang, H. (2021). A review on optimal subsampling methods for massive datasets. *Journal of Data Science*, 151–172. doi: 10.6339/21-JDS999

Yin, K., Mondal, A., Ndeffo-Mbah, M., Banerjee, P., Huang, Q., & Gurarie, D. (2022). Bayesian inference for COVID-19 transmission dynamics in India using a modified SEIR model. *Mathematics*, *10*(21), 4037. doi: 10.3390/math10214037

Zalewski, M. (2002). Ecohydrology—the use of ecological and hydrological processes for sustainable management of water resources / ecohydrologie—la prise en compte de

processus écologiques et hydrologiques pour la gestion durable des ressources en eau. *Hydrological Sciences Journal*, *47*(5), 823–832. doi: 10.1080/02626660209492986

Zhang, J., Vrugt, J. A., Shi, X., Lin, G., Wu, L., & Zeng, L. (2020). Improving simulation efficiency of MCMC for inverse modeling of hydrologic systems with a Kalman-inspired proposal distribution. *Water Resources Research*, *56*(3). doi: 10.1029/2019WR025474

Zhang, J. L. (2014). Comparative investigation of three Bayesian p values. *Computational Statistics & Data Analysis*, *79*, 277–291. doi: 10.1016/j.csda.2014.05.012

Zheng, Y., & Han, F. (2016). Markov Chain Monte Carlo (MCMC) uncertainty analysis for watershed water quality modeling and management. *Stochastic Environmental Research and Risk Assessment*, *30*(1), 293–308. doi: 10.1007/s00477-015-1091-8

Zhibin, Z., Tao, Y., & Li, Z. (2007). Factors affecting hare–lynx dynamics in the classic time series of the Hudson Bay Company, Canada. *Climate Research*, *34*(2), 83–89.