Using replica exchange Hamiltonian Monte Carlo and thermodynamic integration for comparison of dynamic rainfall-runoff models

Damian N. Mingo, Remko Nijzink, Christophe Ley, Stan Schymanski, Jack S. Hale.

April 5, 2023

DTU DRIVEN
Fonds National de la Recherche Luxembourg
UNIVERSITÉ DU LUXEMBOURG
LUXEMBOURG INSTITUTE OF SCIENCE AND TECHNOLOGY LIST

Table of contents

Introduction

Hydrological models

- Hydrology is the study of the movement of water in the environment.
- Hydrologists develop different types of models to understand water dynamics.
- The Hydrologiska Byråns Vattenbalansavdelning (HBV) model and its variants are used in over 50 countries to model hydrological systems (Bergstrom, 2006).
- The HBV model links precipitation with hydrological catchment outflow. The calibrated models are used for flood prediction, water management etc.

Problem statement

- Today, Bayesian inference is widely used in hydrology for parameter identification (Marshall, Nott, & Sharma, 2005).
- However, Bayesian model selection criteria are not widely used, even though the model selection problem is equally important for practitioners.

Why?

1. Computational expense.
2. Poor robustness of algorithms for moderate dimensional problems.
3. Difficulty of implementing those algorithms.
    - Require gradients, not always available.
    - Technical or mathematical expertise.

Main contributions

1. Algorithmic
   - We introduce REHMC+TI, a combination of replica exchange (RE), Hamiltonian Monte Carlo (HMC) and thermodynamic integration (TI) for efficient and robust parameter and marginal likelihood estimation.
2. Statistical
   - We introduce formal posterior predictive checks for ODE-based models.
3. Methodological
   - Algorithms are implemented in the differentiable programming language TensorFlow Probability for flexible future use.

Data

Time series data from Magela Creek Australia

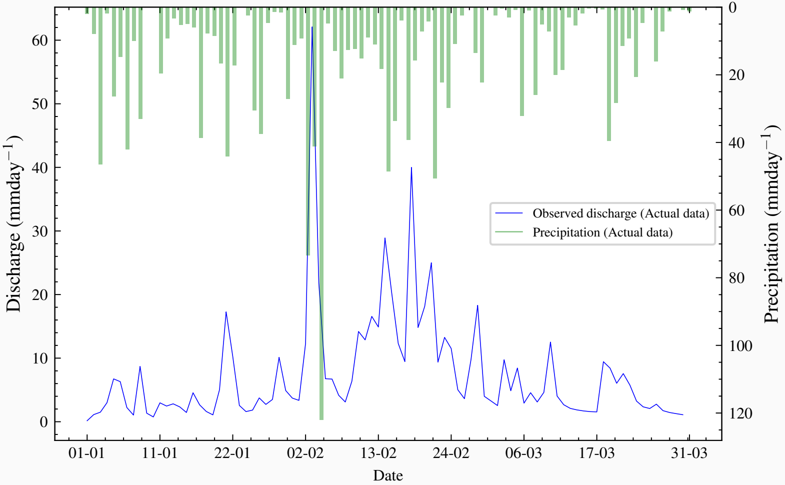


Figure 1: Plot of observed discharge and precipitation from 01-01-1980 to 31-03-1980.

Time series data

- Magela Creek is a gauged catchment.
- The variable of interest is discharge (Q mmday^{-1}).
- The independent variables:
  - Precipitation (mmday^{-1}).
  - Actual evapotranspiration (mmday^{-1}).

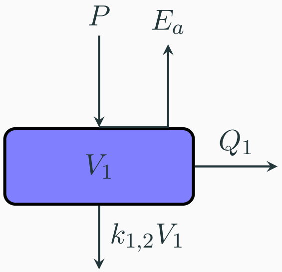
Model development

HBV-type models as an ODE system

We develop a system of ordinary differential equations to mimic the HBV model.

- P **: Precipitation** (L^3T^{-1})
- E_a **: Actual evapotranspiration** (L^3T^{-1})
- Q_1 : Discharge (L^3T^{-1})
- k_1 : outflow recession coefficient (T^{-1})
- $k_{1,2}$: inter-bucket recession coefficient (T^{-1})

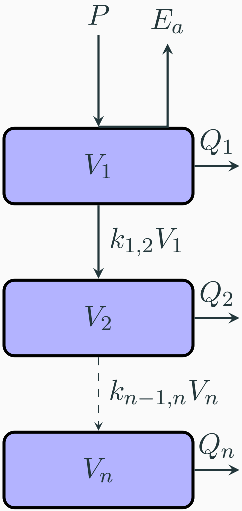
$$\begin{aligned}\frac{\mathrm{d}V_1}{\mathrm{d}t} &= P - E_a - k_{1,2}V_1 - Q_1 \\ &= P - E_a - k_{1,2}V_1 - k_1V_1.\end{aligned}$$



Multi-bucket HBV model

- $V_t := \frac{\mathrm{d}V}{\mathrm{d}t}$ is the derivative of the state with respect to the time variable t.
- $\hat{V} \in \mathbb{R}^n$ are the initial conditions.

$$
\begin{aligned}
(V_1)_t &= P - E_a - k_1 V_1,\\
(V_i)_t &= k_{(i-1)(i)} V_{i-1} - k_i V_i,\ i = 2, \ldots, n-1, n > 2,\\
(V_n)_t &= k_{(n-1)(n)} V_{n-1} - k_n V_n,\\
V(t=0) &= \hat{V},\\
E_a &= \frac{E_p}{V_{\max}} V_1,\\
Q &= \sum_{i=1}^{n} k_i V_i.
\end{aligned}
$$



Likelihood construction

$$y = G_{\text{obs}} G_{\text{sol}}(\theta) + \eta,$$

$$\eta \sim \mathcal{N}(0, \sigma^2 I_p)$$

- $G_{\text{sol}} : \mathbb{R}^q \to X$ maps the parameter vector $\theta \in \mathbb{R}^q$, with $q = 3n$, to the total discharge $Q \in X$ through solving the ODE system.
- G_{obs} : Evaluates the total discharge at specific time points $\{t_1, \ldots, t_p\}$.
- η : noise, assumed Gaussian with covariance $\sigma^2 I_p \in \mathbb{R}^{p \times p}$ with I_p the identity matrix.

$$G := G_{\text{obs}} G_{\text{sol}}$$

$$y | \theta \sim \mathcal{N}(G(\theta), \sigma^2 I_p)$$

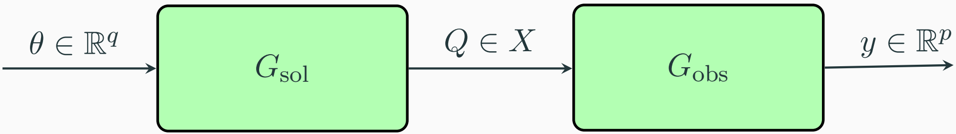


Figure 2: Schematic representation of likelihood construction

Bayesian inference

Theorem (Bayes theorem)

$$\underbrace{p(\theta_n|M_n, y)}_{\text{posterior}} = \frac{\overbrace{p(y|\theta_n, M_n)}^{\text{likelihood}}\,\overbrace{p(\theta_n|M_n)}^{\text{prior}}}{\underbrace{p(y|M_n)}_{\text{marginal (averaged) likelihood}}}$$

$$= \frac{p(y|\theta_n, M_n)p(\theta_n|M_n)}{\int p(y|\theta_n, M_n)p(\theta_n|M_n)d\theta_n}.$$

Bayesian model comparison

- The $\log$ Bayes factor $\log \mathrm{BF}_{ij}$ is obtained by taking the ratio of the $\log$ marginal likelihoods of the i-th and j-th models

$$
\begin{aligned}
\log \mathrm{BF}_{ij} &= \log p(y|M_i) - \log p(y|M_j) \\
&= \log \int p(y|\theta_i, M_i) p(\theta_i|M_i) \, \mathrm{d}\theta_i - \log \int p(y|\theta_j, M_j) p(\theta_j|M_j) \, \mathrm{d}\theta_j .
\end{aligned}
$$

- $\log \mathrm{BF}_{ij} > 1$ is in favour of model i.

Computational aspects

Key techniques

There is usually no analytic solution for the marginal likelihood. Thus, we use sampling-based methods:

1. Thermodynamic integration: Robust method for marginal likelihood estimation that does not require *a priori* choice of bridge/importance distribution.
2. Hamiltonian Monte Carlo: Scales better in high dimensions even when parameters show strong correlations.
3. Replica exchange: Accelerates chain mixing and can handle multimodality, which is inherent in ODE based models.

leading to *Replica Exchange Hamiltonian Monte Carlo* (REHMC).

Thermodynamic integration

We first define the *power posterior* which continuously connects the prior and posterior through the inverse temperature parameter β

$$p(y|\beta) = \int \big[p(y|\theta)\big]^{\beta} \pi(\theta) \, \mathrm{d}\theta, \qquad 0 \leq \beta \leq 1$$

Taking the logarithm and differentiating gives

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

Thermodynamic integration

Further simplifying with the identity $f'(x) = a^x \log a \iff f(x) = a^x$ gives

$$\begin{aligned}\frac{\partial}{\partial\beta}\log p(y|\beta) &= \frac{1}{p(y|\beta)}\int \left[p(y|\theta)\right]^{\beta} \log p(y|\theta)\pi(\theta)\ d\theta \\ &= \int \frac{\left[p(y|\theta)\right]^{\beta}\pi(\theta)}{p(y|\beta)} \log p(y|\theta)\ d\theta \\ &= \mathbb{E}_{p(\theta|y,\beta)}[\log p(y|\theta)].\end{aligned}$$

giving the final result:

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

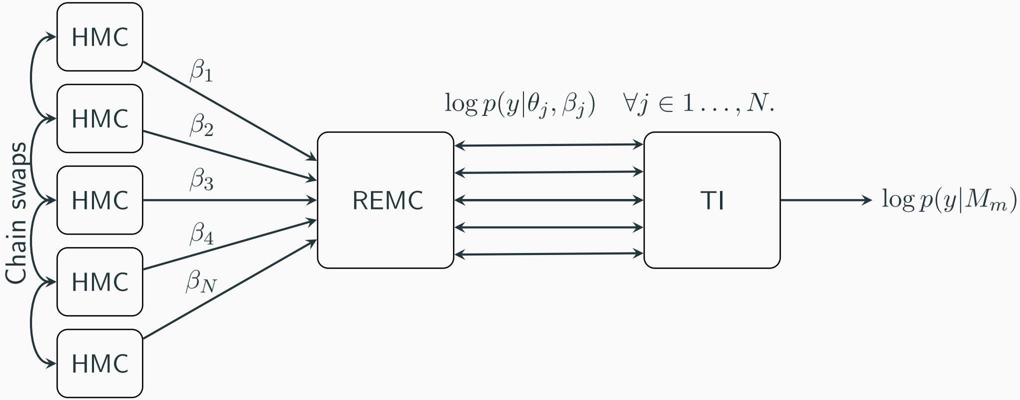
Thermodynamic integration

The trapezoidal + Monte Carlo estimate of the log marginal likelihood can then be written

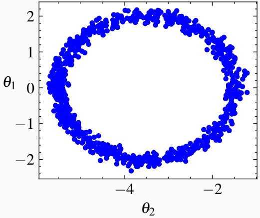
$$\log p(y) \approx \sum_{j=1}^{N} \frac{(\beta_j - \beta_{j-1})}{2} \left[\frac{1}{S} \sum_{i=1}^{S} \log p(y|\theta_i, \beta_j) + \frac{1}{S} \sum_{i=1}^{S} \log p(y|\theta_i, \beta_{j-1}) \right],$$

where N is the number of integration points and S are the number of Monte Carlo samples.

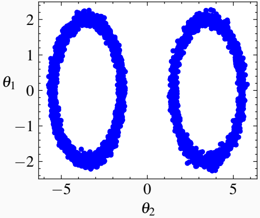
Replica exchange Hamiltonian Monte Carlo (REHMC)



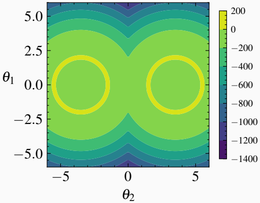
Effectiveness of REHMC



(a) Posterior samples obtained by NUTS.



(b) Samples obtained by REHMC.



(c) Target distribution.

Results

Synthetic example

- Two sets of experiments:
    1. Data generated from the model M_2 with the least number of parameters
    2. Data generated from model M_3 with the highest number of parameters.
- The precipitation and potential evapotranspiration are from the Magala Creek dataset.
- We assigned lognormal priors to all model parameters except $\sigma^2 \sim IG(\alpha, \beta)$.
- The parameters associated with upper buckets are assigned priors with faster timescales (runoff processes vs storage processes).

log marginal likelihood

- Experiment 1, M_2 is the data generating model.

Table 1: log marginal likelihood

M_2	M_3	M_4
201.336	194.722	179.406

- Experiment 2 M_3 is the data generating model.

Table 2: log marginal likelihood

M_2	M_3	M_4
74.815	158.716	152.581

Based on the BF interpretation table by (Kass & Raftery, 1995) we have decisive evidence in favour of the data generating models.

Posterior predictive checks

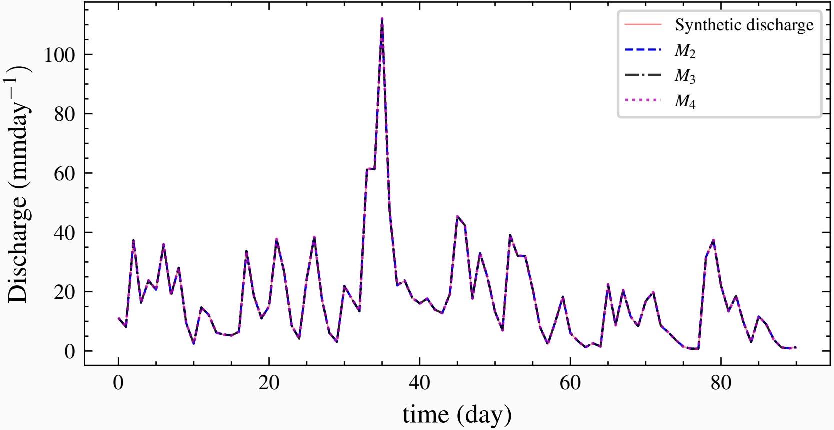


Figure 3: Graphical posterior predictive check

Real discharge data

Results: Real-world data

Table 3: Results using real discharge data

	M_2(95% CI)	M_3(95% CI)	M_4(95% CI)
k_1	1.281(0.893, 1.708)	1.255(0.851, 1.674)	1.298(0.863, 1.752)
k_2	1.506(0.860, 2.180)	1.863(1.142, 2.703)	2.060(1.227, 2.849)
k_3	-	1.342(0.719, 1.997)	1.408(0.775, 2.107
k_4	-	-	1.072(0.574, 1.559)
$k_{1,2}$	1.182(0.788, 1.638)	2.296(0.589, 1.310)	2.292(1.325, 3.327)
$k_{2,3}$	-	0.711(0.481, 0.978)	0.731(0.451, 1.008)
$k_{3,4}$	-	-	0.828(0.541, 1.170)
$\hat{V}_1$	1.066(0.026, 2.776)	1.235(0.027, 3.336)	1.190(0.029, 3.154)
$\hat{V}_2$	0.821(0.061, 1.902)	1.077(0.061, 2.871)	0.997(0.059, 2.557)
$\hat{V}_3$	-	1.220(1.181, 0.029)	1.152(0.029 , 3.228)
$\hat{V}_4$	-	-	1.138(0.029, 3.066)
$V_{\max}$	0.841(0.585, 1.109)	0.941(0.635, 1.251)	0.842(0.595, 1.150)
σ^2	7.591(6.661, 8.787)	7.623(6.620, 8.697)	7.633(6.602, 8.750)
$\log p(y\|M)$	-388.826	**-386.716**	-388.978

Conclusions

1. We have introduced a modelling framework in hydrology that consists of parameter estimation, model selection, and posterior predictive checks.
2. We have illustrated using the gradient-based sampler REHMC that the marginal log-likelihood can be estimated efficiently for ODE-type models.
3. Our framework can be used as an efficient alternative to widely used gradient-free samplers.
4. The entire framework has been implemented in the differentiable programming language TensorFlow Probability.

Thank you

References

Bergstrom, S. (2006). Experience from applications of the hbv hydrological model from the perspective of prediction in ungauged basins. *IAHS publication*, *307*, 97.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, *90*(430), 773–795.

Marshall, L., Nott, D., & Sharma, A. (2005). Hydrological model selection: A bayesian alternative. *Water resources research*, *41*(10).