Bayesian treatment of hydrological models for road rainfall-runoff prediction

Mini-project

Ioannis Zachos

Department of Engineering
University of Cambridge

January 2020

Abstract

[redacted]'s objective to improve the health and safety of its customers is tightly linked to better monitoring of on-road water discharge. Given the increasing climate volatility, a probabilistic approach is adopted to modelling on-road water discharge. This report illustrates the potential of using a hybrid of statistical and hydrological models to better understand on-road water conditions. The study area is a major A-road in [redacted]. The lumped linear reservoir model as well as the spatially-distributed HYMOD model are employed and discharge data is simulated using rainfall and potential evapotranspiration data from the Climate hydrology and ecology research support system (CHESS). The same models are then treated in a Bayesian manner by assigning priors over their tuning parameters. Parameter posteriors and marginal likelihoods are computed using a Sequential Monte Carlo (SMC) sampler. Data fits and Bayes factors are examined to assess goodness of fit and identify the best model, which according to the data is the linear reservoir model.

Keywords: hydrological modelling, Bayesian inference, physics-informed machine learning.

Contents

1 Introduction

The management of key road assets that connect major cities and accommodate millions of vehicles every day is a complicated task. It is increasingly challenging to secure the long-term environmental resilience of that infrastructure, particularly with regards to the effects of erratic rainfall patterns on roads. One way of easing the burden [redacted] faces is to employ sophisticated data-driven hydrological models that monitor the rainfall-runoff (discharge) on the surface of major roads/motorways. This can constitute the basis for a real-time on-road rainfall-runoff monitoring and alert system for road managers and users. However, the increased climate volatility observed due to climate change is polluting the process with noise which can lead to biased estimates of water volumes. In order to make robust inference of the volume of water (otherwise known as water discharge) accumulated on a given road it is necessary to incorporate the uncertainty into the modelling framework.

The approach discussed in this report involves developing a hybrid modelling framework that combines common hydrological models with statistical models to make inference of water discharge on [redacted].

1.1 Literature review

1.1.1 Hydrological models

The vast majority of hydrological modelling studies have been focused on modelling the runoff generation and routing in river, lake or groundwater catchments (or watersheds). Not much attention has been devoted to estimate surface water volume in roads. A study by researchers from the Royal Institute of Technology in Sweden (Kalantari et al. 2014) attempted to use physically-based hydrological models to identify the appropriateness of road drainage structure dimensioning. Another study (Hollis 1988) focused on the hydrological effects of urbanisation. However, neither studies elaborated on any procedural models that can model rainfall-runoff on road surfaces.

A common modelling approach in hydrology is to estimate the (unit) hydrograph, which is the graph of the rate of water flow (discharge) versus time (K. J. Beven 2012, p. 29-33). The rational method (ibid., p. 25-26) is one of the earliest methods for obtaining the peak of a hydrograph as it dates back to the 1851. Due to its poor predictive ability more complicated models have been developed since then. One such model is the Identification of Unit Hydrographs and Component Flows from Rainfall, Evaporation, and Streamflow Data (IHACRES) model (Allen and Liu 2011), which is a conceptual model (i.e. a model based on the conceptions of a hydrologist about the physical processes that govern the modelled system) applied to humid/temperate region catchments. In addition, there are models that have been applied to rivers and bridges, such as the Soil and Water Assessment tool (SWAT) (P. W. Gassman et al. 2007) and the Hydrologic Engineering Center-River Analysis System (HEC-RAS) (US Army Corps of Engineers 2002). The former is being used to assess the impact of land management practices in large watersheds while the latter was used by the United States Army Corps of Engineers in an attempt to manage the rivers, harbors, and other public works under their jurisdiction. The Soil Moisture Accounting and Routing for Transport (SMART) model (Mockler, O'Loughlin, and Bruen 2016) is one of the latest conceptual models developed for agricultural applications. Notable models for impervious land surfaces and streams include the Hydrological Simulation Program FORTRAN (HSPF) (Bicknell et al. 1997), which is a FORTRAN-based program that can simulate the non-point source pollutant loads, hydrology parameters, and aggregate water quality.

Last but not least, two important models are the linear reservoir (LR) (Zeeuw 1973) and the non-linear storage models (NLS) (Botter et al. 2009) are widely applicable lumped models, i.e. models that treat the catchment as a single unit (K. J. Beven 2012, p. 16). Distributed models address the limitation of lumped models as they make predictions in space by discretising the catchment into a finite number of elements (ibid., p. 16). Two of the most prominent distributed models are the probability distributed model (PDM) (Moore and Clarke 1981) and the TOPMODEL (K. Beven 1997). PDM makes use of probability distribution functions to quantify the spatial variability of water storage capacity while TOPMODEL is a topography-based model that can map the predictions back into the catchment to allow additional evaluation of the simulations. Finally, the HYMOD model (Boyle, Gupta, and Sorooshian 2011) is another important models that builds upon the ideas of the PDM.

1.1.2 Uncertainty quantification

Most of the hydrology models explored in the previous section are physics-based parametric models. There is a number of challenges arising in model calibration. First of all, parameter values are often not known *a priori* (K. J. Beven 2012, p. 44). Second of all, the concept of an optimum parameter set may be ill-founded in hydrological modelling (ibid., p. 44), which implies that optimal parameters are not unique. For these reasons recent studies (Vrugt et al. 2003) (Ajami, Duan, and Sorooshian 2007) (Montanari and Brath 2004) have attempted to identify the sources of uncertainty in hydrological models. Novel techniques in hydrology include a hydrological data assimilation approach for estimating model parameters and state variables using particle filters (Moradkhani et al. 2005) and the introduction of a Bayesian framework for model calibration and validation (Kavetski, Franks, and Kuczera 2003). Another notable contribution use of a multimodel Bayesian framework (Ajami, Duan, and Sorooshian 2007) that attempted to distinguish between parameter, input, and structural uncertainties.

Since its introduction in hydrological modelling, Bayesian inference has proved to be a suitable way of managing the different sources of uncertainty in hydrological models. Due to the intractable nature of many physics-based hydrological models, the computation of posterior distributions is usually facilitated by Markov chain Monte Carlo (MCMC) sampling techniques. The use of MCMC as well as variants of Metropolis-Hastings sampling methods has been also been popularised in hydrology (Kuczera and Parent 1998) (Ajami, Duan, and Sorooshian 2007).

1.2 Aims

This projects aims to:

1. Motivate the application of common rainfall-runoff models to roads
2. Identify the potential of using statistical and physics-based (hybrid) approaches to rainfall-runoff modelling
3. Demonstrate the versatility of Bayesian methods of uncertainty quantification in rainfall-runoff modelling

As demonstrated in the proposal brief, this research project is perfectly aligned with HE's key performance indicators (KPIs) with regards to health & safety and road maintenance.

1.3 Scope

Regarding the experimental procedure, two hydrological models (LR and HYMOD) are employed and discharge data is simulated for each for these models. Then, the same models are treated in a Bayesian manner and inference is made on their parameter posteriors based on each simulation dataset, which results in nine models (two model parametrisations for each model structure). This is achieved by assigning prior distributions over the parameters. Posterior inference is made using Sequential Monte Carlo (SMC) sampling and marginal likelihoods are computed as a by-product. Finally, Bayes factors and a 2×2 marginal likelihood matrix are computed for model comparison.

1.4 Study area

The road of interest (Figure 1) was selected from [redacted] dataset which is available on [redacted]. It is a two-way single carriageway A-road located in north-west England whose surface area is approximately 16000 m^2. The road number is A590 and its reference is 210000/NRP/A590/007. According to the 25 m digital elevation model, there are no significant slope variations throughout the road and therefore the road is treated as a flat surface.

In terms of data requirements, the two hydrological models make use of monthly rainfall (in mm/day) and potential evapotranspiration data (in mm/day) from the Climate hydrology and ecology research support system potential evapotranspiration dataset (1961-2012), which is available at $1km$ resolution. Also, [redacted] could not provide on-road road water discharge data (mm^3/day). In order to avoiding treating the road as an ungauged catchment (i.e. a catchment that lacks discharge sensors), discharge data was simulated using the two rainfall-runoff models. Finally, no water drainage data (location and surface area of drainage systems) was made available by [redacted]. This is the case because the [redacted] does not include precise area measurements of drainage systems.

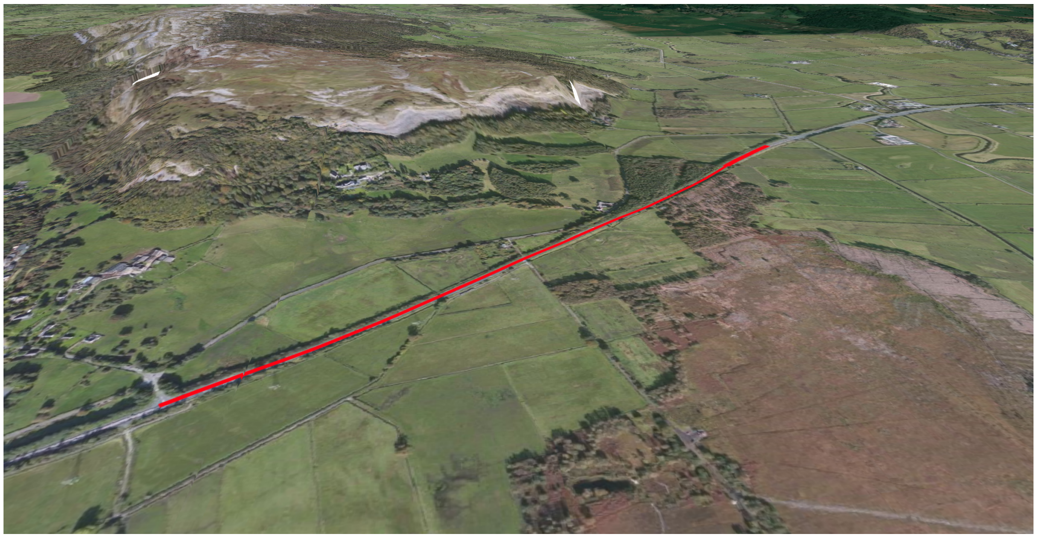


Figure 1: 3D visualisation of road landscape in QGIS. Road is shown in red.

2 Modelling framework

This section outlines the rainfall-runoff models employed to simulate discharge data and make inference on it. In hydrology terms, these models will simulate a hydrograph which will then be mimiced by the mass-balance equations of each model with the aid of statistical machinery.

2.1 Rainfall-runoff models

2.1.1 Linear reservoir model

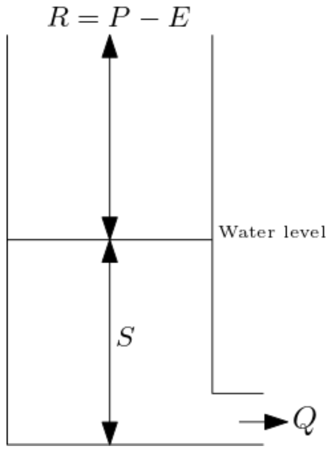


Figure 2: Linear reservoir model pictorial representation.

The linear reservoir model (Zeeuw 1973) shown in Figure 2 is widely used in rainfall-runoff modelling. Water inflows on the surface of the road at any time t are defined as the difference between precipitation $P(t)$ and evapotranspiration $E(t)$, called net rainfall $R(t)$. Note that transpiration due to plants and trees may not be relevant when measuring road surface discharge. However, in this case it will be assumed that water is absorbed by the surrounding flora of the road in Figure 1. Another implicit assumption made here is that water drainage is assumed to be negligible, which is an unrealistic assumption for most A-roads. The lumped nature of the model implies that there is no spatial variability of discharge along the road, which may also be an unrealistic assumption. However, lack of precise road slope data renders a more sophisticated application of the linear reservoir model hard.

The model is assumed to have a constant first-order storage coefficient K (measured in time units), which reflects how quickly the reservoir drains; a smaller value indicates more rapid outflow. It combines continuity and storage-discharge equations, which yields an ordinary differential equation that describes outflow from each reservoir (ibid.). The continuity (or water-balance) equation for tank models is:

$$\frac{dS(t)}{dt} = R(t) - Q(t), \tag{1}$$

where $Q(t)$ is the discharge (i.e. volumetric water flow rate) at time t, $S(t)$ is the water level at time t. Equation 1 indicates that the change in storage over time is the difference between inflows and outflows. The storage-discharge relationship (flow equation) is:

$$Q(t) = \frac{S(t)}{K} \tag{2}$$

Combining equations 1 and 2 yields

$$K\frac{Q(t)}{dt} = R(t) - Q(t), \tag{3}$$

which admits the following solution:

$$Q(t) = R(t) \times (1 - e^{-t/K}).$$

Since discharge $Q(t)$ is a volumetric quantity and is therefore lower-bounded by zero, $R(t) = \max(0, P(t) - E(t))$ holds for the above equation.

Provided the value of K is known, the total hydrograph can be obtained by successively computing the runoff at the each time interval. Otherwise, K can be determined from a data record of rainfall and runoff, which is not available in this application of the model. In the case when the response factor K can be determined from the characteristics of the watershed, the reservoir can be used as a deterministic model. The fact that K is not known *a priori* motivates the need for a Bayesian treatment of the linear reservoir model.

2.1.2 HYMOD model

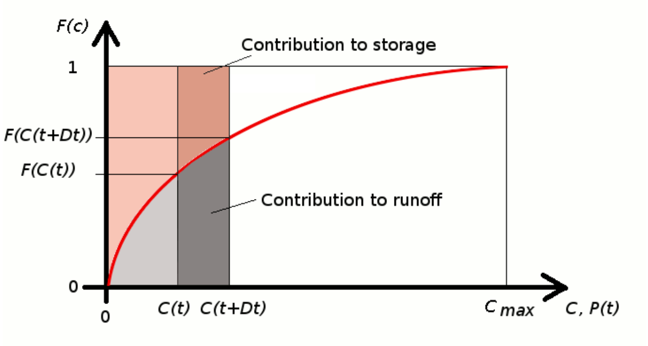


Figure 3: Distribution of soil water storage, surface runoff and stored volume in the HYMOD model (Montanari 2019).

The Hymod model (Boyle, Gupta, and Sorooshian 2011) is increasingly adopted for its capability of providing a good fit in several practical applications. The fundamental assumption it makes is that each point i in the catchment is characterised by a local value of soil water storage C_i, which varies from 0 in the impervious areas up to a maximum value C_{max} in the most permeable location of the catchment. C_i is assumed to be a continuous random variable, so that for an assigned value C_* there is an associated probability for a random location j characterised by C_j less than, or equal to, C_*. Such probability may be interpreted as the fraction $F(C_*)$ of the catchment area where $C_j \leq C_*$. In this case, F is the probability distribution of C_*. The above probability distribution is written as

$$F(C_*) = 1 - \left(1 - \frac{C_*}{C_{max}}\right)^{\beta_k} \tag{4}$$

for $0 \leq C_* \leq C_{max}$.

In the case of impervious road surfaces, it may be intuitive to set $C_i = 0$ for every location i. However, roads have drainage systems in place and the road topography creates local 'storage units'. Therefore, the parameter C_{max} may be interpreted as the maximum storage capacity of say a drainage system on the side of the road. In that way the HYMOD model harnesses the power of a spatially distributed hydrological model and allows for drainage to be taken into account without explicitly using drainage data.

In equation 4, β_k is a parameter which controls the spatial variability of the water storage capacity. It can be verified by numerical simulation that $\beta_k = 0$ implies that the water storage capacity is constant over the catchment and equal to C_{max}; $\beta_k = 1$ implies that water storage is linearly varying from 0 to C_{max}; $\beta_k \to \infty$ implies that water storage tends to the zero, i.e. the road is impervious with no drainage systems in place.

Assuming that a storm event occurs over the road and let $C(t)$ be the time-varying water depth stored in the unsaturated locations of the catchment. Ignoring water losses, such as evapotranspiration, $C(t)$ is equal to the rainfall amount from the beginning of the event. Assuming that the shape of the probability distribution in equation 4, now expressed in terms of $C(t)$, is the one reported in Figure 3, it can be easily shown that the water volume stored in the catchment at time t is given by

$$W(t) = C(t) - \left[\int_0^{C(t)} F(x)dx\right] \tag{5}$$

The integral on the right hand side of the above equation is the area below the red line in Figure 3. Each area increment is given by the product of rainfall at each time step and the fraction $F(C(t))$ of saturated area at that time, which is equal to the surface runoff. Conversely, the area above the curve gives the global storage $W(t)$ into the catchment as a weighted average of $C(t)$. The movement of surface runoff and water storage is depicted in Figure 4. After saturation, the storage in the catchment reaches a plateau and the contribution of surface runoff is given by the excess rainfall, which is the fraction of saturated area.

Evaluating the integral in equation 5 results in the water volume stored in the catchment being equal to

$$W(t) = \frac{C_{max}}{\beta_k + 1}\left[1 - \left[1 - \frac{C(t)}{C_{max}}\right]^{\beta_k+1}\right] \tag{6}$$

Inverting the above equation yields

$$C(t) = C_{max}\left[1 - \left(1 - W(t)\frac{\beta_k + 1}{C_{max}}\right)^{\frac{1}{\beta_k+1}}\right] \tag{7}$$

An upper bound for W_{max} can be derived by setting $C(t) = C_{max}$ $\forall t \in \{1, \dots, T\}$:

$$W_{max} = \frac{C_{max}}{\beta_k + 1}$$

The above equations allow for an easy application of the Hymod model through a numerical simulation, that is usually carried out by adopting a time step Δt that is equal to observational time step of rainfall and water flow (one day in this case). At any given time t, the value of $C(t)$ is known to be equal to the cumulative rainfall depth from the beginning of the event at time t. Therefore, $W(t)$ can be easily computed as well by using the above relationships. At time $t+1$, $C(t+1) = C(t) + P(t)$, where $P(t)$ is rainfall, under the conditions that $C(t+1) = C_{max}$ and $C(t) + P(t) > C_{max}$. Therefore, a first contribution to surface runoff can be computed through the relationship $ER_1(t) = \max(C(t) + P(t) - C_{max}, 0)$.

A second contribution to the surface runoff is made by the water volume that cannot be absorbed by the catchment because part of the catchment area got saturated in the last time step. This contribution is given by $ER_2(t) = (C(t+1) - C(t)) - (W(t+1) - W(t))$. The total contribution to the surface runoff during $[t, t+1]$ is therefore given by the sum of $ER_1(t)$ and $ER_2(t)$.

At this stage, the water losses are computed through the relationship

$$E(t) = \left(1 - \frac{\frac{C_{max}}{\beta_k+1} - W(t)}{\frac{C_{max}}{\beta_k+1}}\right) E_p(t), \tag{8}$$

where $E_p(t)$ is the potential evapotranspiration at time t provided in the raw data. Then, the water storage at time $t+1$ is given by $W(t+1) = W(t) - E_p(t)$. Note the actual evapotranspiration is subtracted from the stored water volume after $ER_1(t)$ and $ER_2(t)$ are computed.

The total contribution $ER(t) = ER_1(t) + ER_2(t)$ to the surface runoff is then divided into 2 components: $\alpha ER_2(t) + ER_1(t)$ which represent the fast flow and $(1-\alpha)ER_2(t)$ which is the slow flow. Fast flow is propagated through a cascade (series) of linear reservoirs ($n_{reservoirs}$ of them) with the same constant coefficient k_q, while slow flow is instead propagated through a single linear reservoir with parameter k_s. Let $Q_f(t)$ and $Q_s(t)$ be the discharges from the fast and slow flows, respectively. Then, the total discharge is equal to $Q(t) = Q_f(t) + Q_s(t)$.

The process above is then repeated for each time step up to time T and therefore the computational complexity of simulating a discharge time series is $\mathcal{O}(T)$ given that parameters C_{max}, β_k, α, k_q and k_s are known.

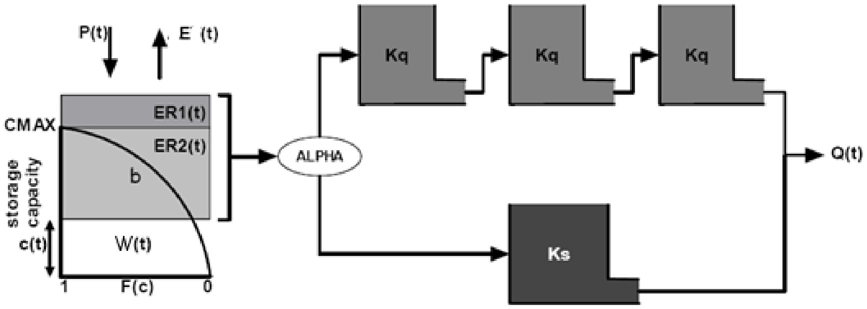


Figure 4: A schematic representation of the Hymod model.

2.1.3 Simulations

The models outlined in the two previous sections were used to simulate discharge data based on the raw rainfall and evapotransporitation data. The simulations were coded in Python using the `scipy.integrate` package. To avoid getting nonsensical discharge values, net rainfall (rainfall minus potential evapotranspiration) was lower bounded by zero. Tables 1 and 2 show the true parameter values for the LR and HYMOD models. True parameters were chosen based on literature (Moore and Clarke 1981). The boundary conditions $Q(0)$ and $W(0)$ were assumed to be known a priori and were not tuned. Figure 5 shows the resulting simulated monthly discharge values for both models as well as the net rainfall patterns. It seems that despite the nature of the models (lumped versus distributed) and the model complexities (two versus six tuning parameters) the two simulations are highly correlated. This is expected as both models leverage the same rainfall and evapotranspiration data.

Parameter	Value
$Q(0)$	0.01
k	0.8
σ	0.5

Table 1: Linear reservoir model simulation parameters.

Parameter	Value
$Q(0)$	0.01
$W(0)$	0.01
C_{max}	300
β_k	0.3
α	0.4
k_{slow}	1.5
k_{fast}	1.1
$n_{reservoirs}$	3
σ	0.5

Table 2: HYMOD model simulation parameters.

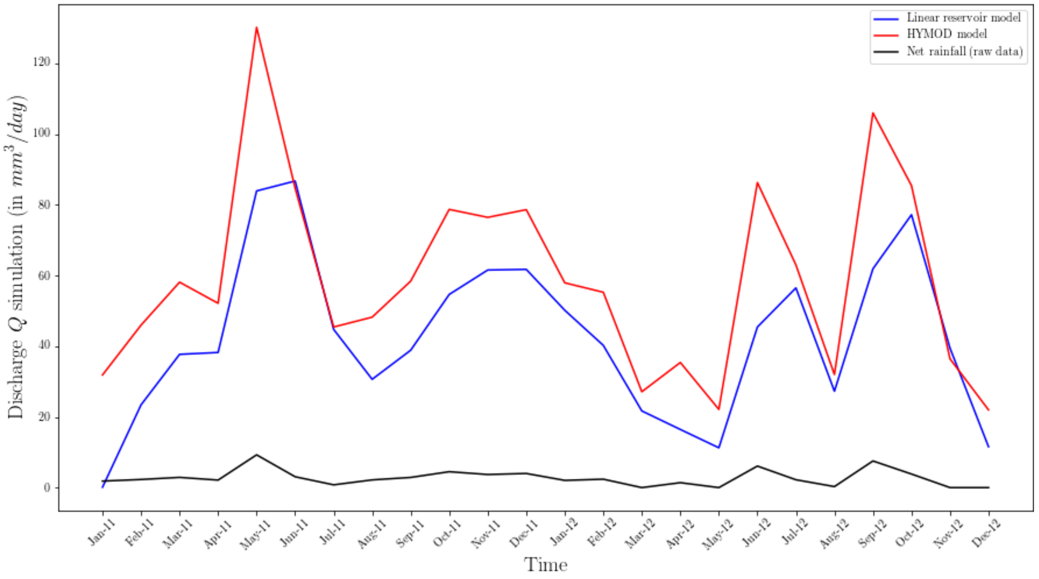


Figure 5: Monthly discharge simulations based on LR and HYMOD models shown in conjunction with net rainfall patterns.

2.2 Bayesian inference

Let a typical hydrological model M be represented as follows:

$$\boldsymbol{Y} = M(\tilde{\boldsymbol{X}}, \boldsymbol{\theta}),$$

where $\boldsymbol{Y} \in \mathbb{R}^{T \times N}$ represents the response matrix of the catchment (i.e. discharge). The (non-)linear hydrological model is denoted by $M(\cdot)$ while $\boldsymbol{\theta} \in \mathbb{R}^p$ corresponds to the set of model parameters, and $\tilde{\boldsymbol{X}}$ the matrix of observed (input) covariates (rainfall and evapotranspiration). In practise, the $\boldsymbol{Y}$ process is noisy and is therefore represented with additive residuals as follows:

$$\tilde{\boldsymbol{Y}} = M(\tilde{\boldsymbol{X}}, \boldsymbol{\theta}) + \epsilon(\boldsymbol{\theta}),$$

where $\tilde{\boldsymbol{Y}}$ is the set of noisy experimental observations observed at T experimental time points for the N states (for both LR and HYMOD models $N = 1$). The additive noise $\epsilon(\boldsymbol{\theta})$ in the process is independent and identically distributed Gaussian noise with mean zero and constant unknown variance σ^2.

In a Bayesian framework $\boldsymbol{\theta}$ is assumed to be a random variable with an associated probability distribution. *A priori* to observing any discharge data, the beliefs about the parameter values are expressed in the prior distribution $\pi(\boldsymbol{\theta})$. Therefore, by Bayes' rule

$$p(\boldsymbol{\theta}|\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}) = \frac{p(\tilde{\boldsymbol{Y}}|\tilde{\boldsymbol{X}}, \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int p(\tilde{\boldsymbol{Y}}|\tilde{\boldsymbol{X}}, \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}},$$

where $p(\tilde{\boldsymbol{Y}}|\tilde{\boldsymbol{X}}, \boldsymbol{\theta})$ is referred to the likelihood of the response given the covariates and parameters, $p(\boldsymbol{\theta}|\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}})$ is the parameter posterior distribution and $\int p(\tilde{\boldsymbol{Y}}|\tilde{\boldsymbol{X}}, \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ is the marginal likelihood or evidence of model M (also denoted as $p(\tilde{\boldsymbol{Y}}|M)$).

Due to the intractability of the parameter posterior and marginal likelihood, Sequential Monte Carlo (SMC) was employed (Kantas et al. n.d.). Compared to other sampling techniques, SMC facilitates the computation of the marginal likelihood with some extra computational cost. By selecting an initial population of particles that sample from the prior and estimate the posterior distributions, the SMC sampler will recursively update that population based on its posterior mass to generate good posterior estimates. Given the nature of this particular application, there is no need for a convergence criterion as the true model parameters are known.

The linear reservoir model has two tuning parameters: k and σ. The priors over these two parameters are

$$k \sim Uniform(0.01, 5)$$
$$\sigma \sim Gamma(2, 4).$$

The HYMOD model has six tuning parameters: $c_{max}, \alpha, \beta_k, k_{slow}, k_{fast}$ and σ. The priors over these six parameters are

$$\begin{aligned}
c_{max} &\sim Uniform(1, 400) \\
\alpha &\sim Beta(2, 3) \\
\beta_k &\sim Gamma(2, 6) \\
k_{slow} &\sim Uniform(0.01, 5) \\
k_{fast} &\sim Uniform(0.01, 2) \\
\sigma &\sim Gamma(2, 4).
\end{aligned}$$

For both models the likelihood of the data is assumed to be

$$p(\tilde{\boldsymbol{Y}}|\tilde{\boldsymbol{X}}, \boldsymbol{\theta}) \sim Normal(X(\tilde{\boldsymbol{X}}, \boldsymbol{\theta}), \sigma^2),$$

where $X(\tilde{\boldsymbol{X}}, \boldsymbol{\theta})$ is the ODE solution to the $\tilde{\boldsymbol{X}}$ given the choice of $\boldsymbol{\theta}$.

For model comparison purposes, the Bayes factor is going to be computed. Assuming the prior over models is uniform, the Bayes factor is equal to the likelihood ratio

$$\frac{P(\tilde{\boldsymbol{Y}}|M_{LR})}{P(\tilde{\boldsymbol{Y}}|M_{HYMOD})} = \frac{P(M_{LR}|\tilde{\boldsymbol{Y}})}{P(M_{HYMOD}|\tilde{\boldsymbol{Y}})},$$

which is equal to the ratio of marginal likelihoods. If the ratio is greater than one, the evidence supports that the model in the numerator is better than model in the denominator.

3 Experimental results

This section illustrates the results of training the two models on two simulated datasets. Figures 7 and 8 depict the estimated parameter posterior distribution of the LR and HYMOD models trained on their simulated datasets.

3.1 Linear reservoir model

According to Figure 7, the posterior of k is approximately Gaussian while the σ posterior is closer to a Gamma distribution (which is also the prior on σ). The posterior sample mean is very close to the true value for both values, which is indicative of convergence. The posterior predictive mean and 95% interval shown in Figure 6 illustrates a good data fit. It is clear that the LR model is more uncertain at the peaks of discharge data than it is at the troughs. This is because the distribution of Q is positively skewed.

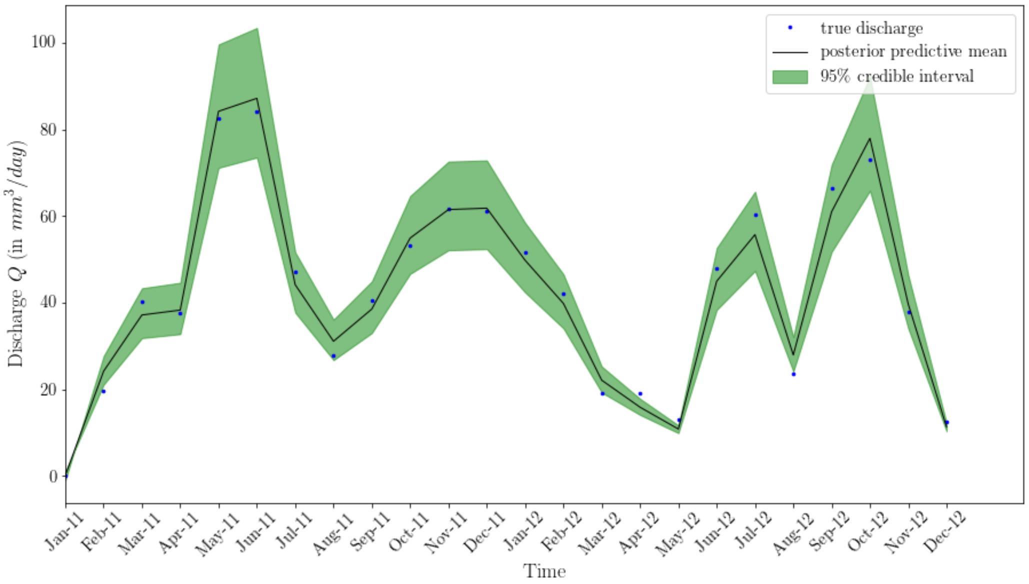
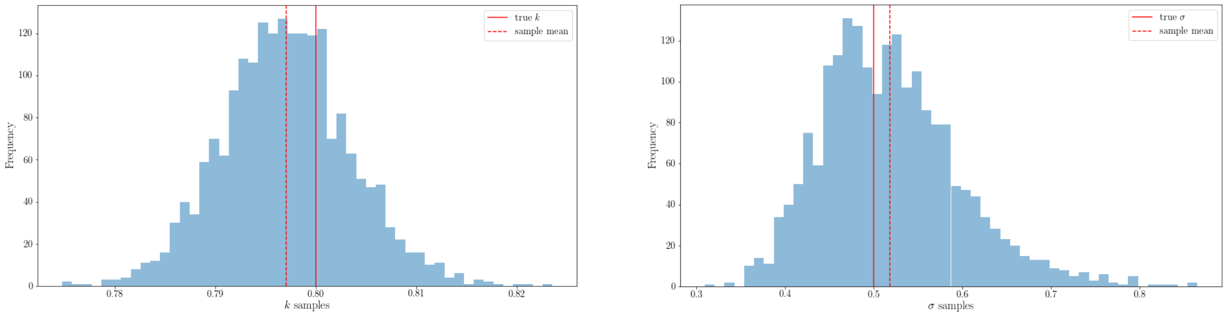


Figure 6: Linear reservoir model posterior predictive distribution of monthly discharge simulations.



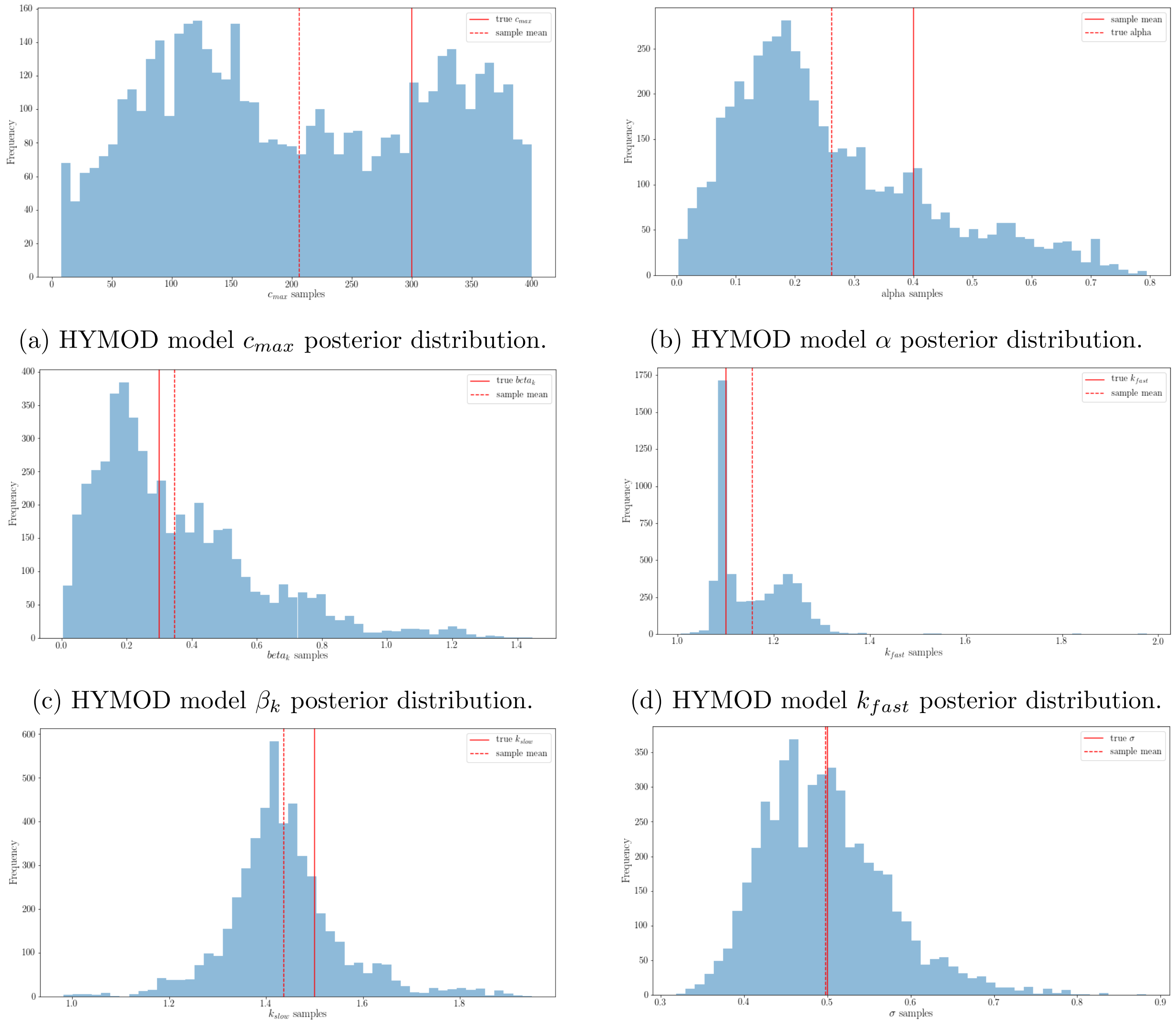
(a) LR model k posterior distribution.

(b) LR model σ posterior distribution.

Figure 7: Linear reservoir model parameter posterior distributions.

3.2 HYMOD model

The increased complexity of the HYMOD model (seven tuning parameters) necessitated a larger particle population size for convergence of the posteriors to be achieved. Figures 8e and 8d show that the posterior means of β_k,k_{fast},k_{slow}, and σ are very close to their true values. The posterior distributions of k_{slow}, and σ are approximately Gaussian and Gamma, respectively. Also, the distributions of β_k and k_{fast} resemble multimodal-like distributions from the exponential family. Last but not least, the distributions of c_{max} and k_{fast} seem to be bimodal. The multimodality of these distributions may be attributed to what hydrologists refer to as *equifinality* (K. J. Beven 2012, p. 44-45). This concept captures the idea that there may be more than one equally valid parameter sets and model structures.



(a) HYMOD model c_{max} posterior distribution.

(b) HYMOD model α posterior distribution.

(c) HYMOD model β_k posterior distribution.

(d) HYMOD model k_{fast} posterior distribution.

(e) HYMOD model k_{slow} posterior distribution.

(f) HYMOD model σ posterior distribution.

Figure 8: HYMOD model parameter posterior distributions.

Also, the fact that the number of data points is 24 intuitively provides room for multiple ‘optimal’ parameter configurations. Therefore, posterior convergence is also achieved for the HYMOD model. This belief is reinforced by the goodness of fit shown in Figure 9.

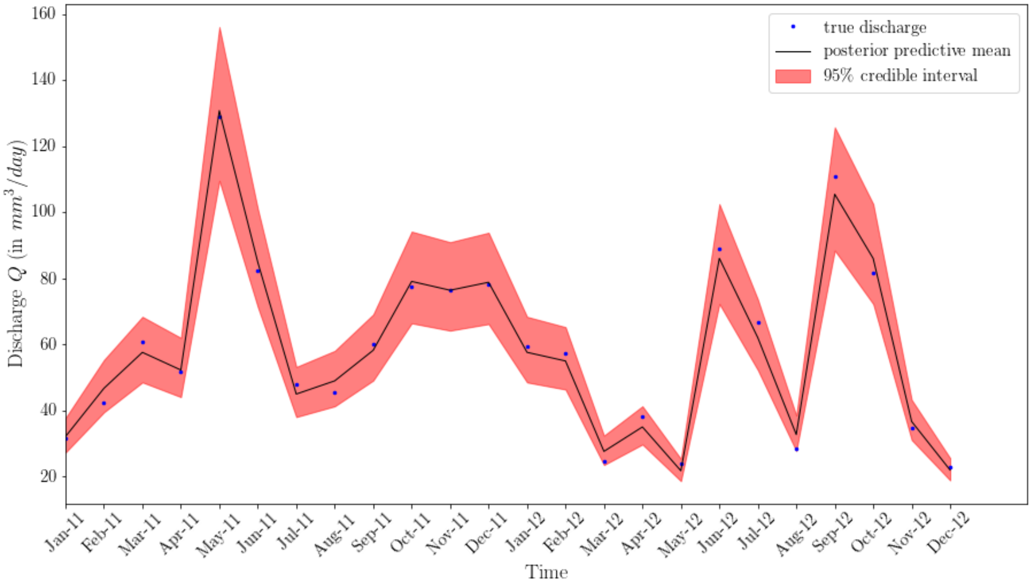


Figure 9: HYMOD model posterior predictive distribution of monthly discharge simulations.

3.3 Model comparison

The two models appear to have a similar data fit as shown in Figures 6 and 9. The marginal likelihoods of the LR model is however higher than the HYMOD model's according to Table 3a. Thus, the additional complexity of the HYMOD model did not provide a better explanation of the data. The Bayes factor of 1.3 implies that there is sufficient evidence that the LR model is better than the HYMOD model when they are both applied on data generated from the same model. As a result, the lumped hydrological model appears to be more powerful than the spatially-distributed model when tested against the simulated discharge datasets. In the two cases when one of the two models is misspecified it is evident that the misspecified model is penalised as expected (see Bayes factors along the diagonal of Table 3b). The Bayes factor of 5.0 implies that the LR model is better at describing HYMOD-simulated data than the HYMOD model is at modelling LR-simulated data. Hence, the evidence suggests that the LR model is slightly more versatile than the HYMOD model.

4 Conclusions and recommendations

[redacted]'s management of critical road infrastructure is a demanding task that necessitates the use of an evidence-based approach. This report proposes a data-driven approach of assessing the impact of rainfall on the Strategic Road Network. The suggested modelling approach of estimating on-road water accumulation provides a basis for future on-site monitoring tools that will help

		model	
		LR	HYMOD
data	LR	1.2	4×10^{-4}
	HYMOD	2×10^{-3}	0.9

(a) Marginal likelihoods of LR and HYMOD models trained on simulation data.

		j	
		LR	HYMOD
i	LR	3000	5.0
	HYMOD	1.3	0.002

(b) Bayes factors $\frac{P(\tilde{\mathbf{Y}}_j|M_{LR})}{P(\tilde{\mathbf{Y}}_i|M_{HYMOD})}$ for LR and HYMOD models.

Table 3: Model comparison of LR and HYMOD models.

[redacted] better maintain its infrastructure while also providing better signal about road conditions to its customers. By harnessing the power of statistical and hydrological models, many of the limitations encountered in traditional hydrological modelling are addressed, such as parameter uncertainty. The probabilistic approach adopted allows [redacted] to make robust inference of water discharge on its roads. The two different types of models employed (lumped and distributed) depict two common approaches in traditional hydrological modelling. The estimated posterior distributions of model parameters indicate convergence of the particles to the true posterior for both the LR and HYMOD models. However, the convergence is more apparent for the LR model. For the HYMOD model, the significant discrepancies observed between the sample mean and true values of two model parameters can be attributed to *equifinality* of hydrological models. Given the small dataset size and the goodness of fit of the HYMOD model as evidenced by the posterior predictive and the marginal likelihood, it can be argued that convergence was also achieved for the HYMOD model. In terms of model comparison, the Bayes factors indicate that the LR is a more powerful and versatile model than the HYMOD model as the additional complexity of the HYMOD model does not explain more aspects of the catchment data.

The existing modelling framework can be improved by using a larger dataset over a wider study area (i.e. more than one roads) to verify the findings derived from this research. A holistic approach would also require real discharge data and precise drainage data. More advanced hydrological models with more climatic and GIS features can also be employed to assess the effect of different model structures on predictive performance. Last but not least, catchment characteristics are not often best captured by a single hydrological model. For that reason, this framework can be extended to incorporate multiple Bayesian models when making inference.

References

Ajami, Newsha K., Qingyun Duan, and Soroosh Sorooshian (Jan. 2007). "An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction". In: *Water Resources Research* 43.1. ISSN: 00431397. DOI: 10.1029/2005WR004745. URL: http://doi.wiley.com/10.1029/2005WR004745.

Allen, Gerald R. and Guangdong Liu (2011). "IHACRES Classic: Software for the Identification of Unit Hydrographs and Component Flows". In: *Groundwater* 49.3, pp. 305–308. DOI: 10.1111/j.1745-6584.2011.00814.x. eprint: https://ngwa.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1745-6584.2011.00814.x. URL: https://ngwa.onlinelibrary.wiley.com/doi/abs/10.1111/j.1745-6584.2011.00814.x.

Beven, K. J. (2012). *Rainfall-runoff modelling: the primer.* Wiley-Blackwell.

Beven, Keith (1997). "TOPMODEL: a critique". In: *Hydrological processes* 11.9, pp. 1069–1085.

Bicknell, B. R. et al. (1997). *Hydrological Simulation Program FORTRAN.* URL: https://ntrl.ntis.gov/NTRL/dashboard/searchResults/titleDetail/PB97193114.xhtml (visited on 12/16/2019).

Botter, Gianluca et al. (Oct. 2009). "Nonlinear storage-discharge relations and catchment streamflow regimes". In: *Water Resources Research* 45.10. ISSN: 00431397. DOI: 10.1029/2008WR007658. URL: http://doi.wiley.com/10.1029/2008WR007658.

Boyle, Douglas P., Hoshin V. Gupta, and Soroosh Sorooshian (Nov. 2011). "Multicriteria calibration of hydrologic models". In: pp. 185–196. DOI: 10.1029/ws006p0185.

Hollis, G E (Oct. 1988). "Rain, Roads, Roofs and Runoff: Hydrology in Cities". In: *Geography* 73.1, pp. 9–18. ISSN: 0016-7487. URL: https://www.jstor.org/stable/40571337%20http://files/5/Hollis%20-%201988%20-%20Rain,%20Roads,%20Roofs%20and%20Runoff%20Hydrology%20in%20Cities.pdf.

Kalantari, Zahra et al. (Mar. 2014). "On the utilization of hydrological modelling for road drainage design under climate and land use change". In: *Science of the Total Environment* 475, pp. 97–103. ISSN: 18791026. DOI: 10.1016/j.scitotenv.2013.12.114.

Kantas, N et al. (n.d.). *An Overview of Sequential Monte Carlo Methods for Parameter Estimation in General State-Space Models.* Tech. rep.

Kavetski, Dmitri, Stewart W. Franks, and George Kuczera (2003). "Confronting Input Uncertainty in Environmental Modelling". In: *Calibration of Watershed Models.* American Geophysical Union (AGU), pp. 49–68. ISBN: 9781118665671. DOI: 10.1029/WS006p0049. eprint: https://agupubs.

onlinelibrary.wiley.com/doi/pdf/10.1029/WS006p0049. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/WS006p0049.

Kuczera, George and Eric Parent (Nov. 1998). "Monte Carlo assessment of parameter uncertainty in conceptual catchment models: the Metropolis algorithm". In: *Journal of Hydrology* 211.1, pp. 69–85. DOI: 10.1016/S0022-1694(98)00198-X.

Mockler, Eva, Fiachra O'Loughlin, and Michael Bruen (May 2016). "Understanding hydrological flow paths in conceptual catchment models using uncertainty and sensitivity analysis". In: *Computers and Geosciences* 90, pp. 66–77. ISSN: 00983004. DOI: 10.1016/j.cageo.2015.08.015.

Montanari, Alberto (2019). *Rainfall-Runoff modeling.* URL: https://distart119.ing.unibo.it/albertonew/?q=node/58 (visited on 11/14/2019).

Montanari, Alberto and Armando Brath (2004). "A stochastic approach for assessing the uncertainty of rainfall-runoff simulations". In: *Water Resources Research* 40.1. DOI: 10.1029/2003WR002540. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2003WR002540. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2003WR002540.

Moore, R. J. and R. T. Clarke (1981). "A distribution function approach to rainfall runoff modeling". In: *Water Resources Research* 17.5, pp. 1367–1382. DOI: 10.1029/WR017i005p01367. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/WR017i005p01367. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/WR017i005p01367.

Moradkhani, H. et al. (2005). "Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter". In: *Water Resources Research* 41. DOI: 10.1029/2004WR003604.

P. W. Gassman et al. (2007). "The Soil and Water Assessment Tool: Historical Development, Applications, and Future Research Directions". In: *Transactions of the ASABE* 50.4, pp. 1211–1250. ISSN: 2151-0040. DOI: 10.13031/2013.23637. URL: http://elibrary.asabe.org/abstract.asp??JID=3%7B%5C&%7DAID=23637%7B%5C&%7DCID=t2007%7B%5C&%7Dv=50%7B%5C&%7Di=4%7B%5C&%7DT=1.

US Army Corps of Engineers (2002). *HEC-RAS River Analysis System.* Tech. rep. Hydrologic Engineering Center. URL: www.hec.usace.army.mil.

Vrugt, Jasper A. et al. (2003). "A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters". In: *Water Resources Research* 39.8. DOI: 10.1029/2002WR001642. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2002WR001642. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2002WR001642.

Zeeuw, J. W. de (1973). "Hydrograph analysis for areas with mainly groundwater runoff". In: *In: Drainage Principle and Applications* 16, pp. 321–358.

Appendices

A Linear reservoir model Bayesian inference

```
import theano
from theano import *
import theano.tensor as tt
from theano.compile.ops import as_op
from models.LinearReservoirModel import LinearReservoirModel as LRM
from tqdm import tqdm
import pandas as pd
import numpy as np
import pymc3 as pm
import json

''' Import simulated data '''

# ...

''' Compute posterior samples '''

# Instantiate linear reservoir statistical model
lrm = LRM(nr,true_args)
@as_op(itypes=[tt.dscalar], otypes=[tt.dmatrix])
def th_forward_model(param1):
    parameter_list = [param1]
    th_states = lrm.simulate(parameter_list,true_args.fatconv)
    return th_states

# Initialise dataframe to store parameter posteriors
# Loop over simulated datasets and compute marginal
for mi in tqdm(model_discharges.keys()):
    print(f'LRM posterior sample generation using {mi} data')
    with pm.Model() as LR_model:

        # Priors for unknown model parameters
        k = pm.Uniform('k', lower=0.01, upper=args.kmax)

        # Priors for initial conditions and noise level
        sigma = pm.Gamma('sigma',alpha=args.alpha,beta=args.beta)

        # Compute forward model
        forward = th_forward_model(k)

        # Compute likelihood
        Q_obs = pm.Normal('Q_obs', mu=forward, sigma=sigma, observed=model_discharges[mi])

        # Fix random seed
        np.random.seed(args.randomseed)

        # Initial points for each of the chains
        startsmc = [{'k':np.random.uniform(0.01,args.kmax,1)} for _ in range(args.nchains)]

        # Sample posterior
        trace_LR = pm.sample(args.nsamples, progressbar=True, start=startsmc, step=pm.SMC())

        # Compute marginal likelihood
        ml = LR_model.marginal_likelihood #-np.log(LR_model.marginal_likelihood)
        print('Marginal Likelihood:',ml)
```

B HYMOD model Bayesian inference

```
import theano
from theano import *
import theano.tensor as tt
from theano.compile.ops import as_op
from models.HymodModel import HymodModel as HYMOD
from tqdm import tqdm
import pandas as pd
import numpy as np
import pymc3 as pm
import json

''' Import simulated data '''

# ...

''' Compute posterior samples '''

# Instantiate linear reservoir statistical model
hymod = HYMOD(rn, et, true_args)
@as_op(itypes=[tt.dscalar, tt.dscalar, tt.dscalar, tt.dscalar, tt.dscalar], otypes=[tt.dmatrix])
def th_forward_model(param1,param2,param3,param4,param5):
    parameter_list = [param1,param2,param3,param4,param5]
    th_states = hymod.simulate(parameter_list, true_args)
    return th_states

# Loop over simulated datasets and compute marginal
for mi in tqdm(model_discharges.keys()):
    with pm.Model() as HYMOD_model:
        # Priors for unknown model parameters
        cmax = pm.Uniform('cmax',lower=1.0,upper=args.c_max,transform=None)
        kfast = pm.Uniform('kfast',lower=0.01,upper=args.kfast_max,transform=None)
        kslow = pm.Uniform('kslow',lower=0.01,upper=args.kslow_max,transform=None)
        betak = pm.Gamma('betak', alpha=args.betak_alpha,beta=args.betak_beta,transform=None)
        alfa = pm.Beta('alfa',alpha=args.alfa_alpha,beta=args.alfa_beta,transform=None)

        # Priors for initial conditions and noise level
        sigma = pm.Gamma('sigma',alpha=args.sigma_alpha,beta=args.sigma_beta)

        # Compute forward model
        forward = th_forward_model(cmax,betak,alfa,kfast,kslow)

        # Compute likelihood
        Q_obs = pm.Normal('Q_obs', mu=forward, sigma=sigma, observed=model_discharges[mi])

        # Sample posterior
        trace_HYMOD = pm.sample(args.nsamples, progressbar=True, step=pm.SMC(),random_seed=args.randomseed)

        # Compute negative marginal likelihood
        ml = HYMOD_model.marginal_likelihood #-np.log(HYMOD_model.marginal_likelihood)
        print('Marginal Likelihood:',ml)
```