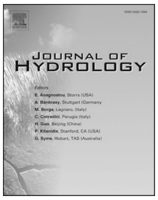
Journal of Hydrology 612 (2022) 128057

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol



Research papers

Application of stochastic time dependent parameters to improve the characterization of uncertainty in conceptual hydrological models

Marco Bacci [a,*], Marco Dal Molin [a,b], Fabrizio Fenicia [a], Peter Reichert [a,c], Jonas Šukys [a]

[a] *Eawag: Swiss Federal Institute of Aquatic Science and Technology, Ueberlandstrasse 133, 8600 Dübendorf, Switzerland*
[b] *Centre of Hydrogeology and Geothermics (CHYN), University of Neuchâtel, Neuchâtel, Switzerland*
[c] *Department of Environmental Systems Science, ETH Zurich, Zurich, Switzerland*

ARTICLE INFO

This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Shailesh Kumar Singh, Associate Editor

Keywords:
Stochastic parameters
Uncertainty quantification
Conceptual models
Bayesian inference
Particle filter
Cross-validation

ABSTRACT

The traditional description of a hydrological system with a deterministic, conceptual model and a lumped output error term does not explicitly consider the main mechanisms of uncertainty generation due to approximate process representation, unobserved variability in processes and influence factors, and input uncertainty. In this study, we test the description of such intrinsic uncertainty in conceptual models by making process rates stochastic through stochastic, time-dependent rate parameters. We analyze the advantages and challenges of this approach by using Bayesian inference to jointly estimate model parameters, parameters of the stochastic processes, and the time series of the stochastic parameters. Numerically, we use a particle filter to infer the stochastic time series and to approximate the marginal likelihood for Markov chain sampling of the constant parameters. Compared to the lumped error formulation, we achieve a more realistic description of uncertainty, as we obtain larger errors in intrinsic states, larger uncertainty during prediction than calibration periods (a feature missing for simple lumped error models), and autocorrelated model outputs. However, the additional degrees of freedom introduced with stochastic parameters can lead to an unintentional compensation for model deficits or input errors. This problem is symptomatic of model structure inadequacy or poor selection of the parameters that are made stochastic, and can be diagnosed through cross-validation and careful posterior analysis. The proposed approach is computationally more demanding and its implementation more challenging than the traditional description of hydrological systems using a deterministic model with a lumped error term. However, its advantages both in providing a more realistic representation of uncertainties and in diagnosing model deficiencies suggest its adoption and further development in future studies.

1. Introduction

Process based catchment models are traditionally deterministic, and their uncertainty is accounted for by separated, lumped error terms, which are applied to their output (e.g., McInerney et al., 2017). Such lumped error models describe the effect of all sources of uncertainty, including input, model structure, intrinsic and externally caused randomness, and observation error on modelled output (e.g., Schoups and Vrugt, 2010; Reichert and Schuwirth, 2012; Evin et al., 2013; Evin et al., 2014; McInerney et al., 2017; Ammann et al., 2019). This approach leads to an efficient inference procedure at the cost of the need for an empirical error model parameterization and of the underestimation of the uncertainty of all internal model states (Reichert et al., 2021).

To mitigate these problems, it has been argued in favor of modeling the uncertainty where it supposedly arises, rather than grouping all contributions in a single error term on the output of the model. This can be done by making the hydrological process model stochastic (e.g., Moradkhani et al., 2005; Kuczera et al., 2006; Liu and Gupta, 2007; Reichert and Mieleitner, 2009; Reichert et al., 2021). The first arguments for the need of stochastic models in hydrology were introduced more than 50 years ago by Mandelbrot and Wallis (1968). More recently, they consider general reasoning on the lack of suitability of deterministic models to describe hydrological features (Kuczera et al., 2006), the intrinsically variable nature of hydrological parameters (Liu and Gupta, 2007), the characterization of structural model deficits (Leisenring and Moradkhani, 2011), and the opportunities for their identification (Wagener et al., 2003; Reichert et al., 2021). Within the domain of stochastic approaches to hydrology, stochastic time-

[*] Corresponding author.
E-mail address: marco.bacci@eawag.ch (M. Bacci).

https://doi.org/10.1016/j.jhydrol.2022.128057
Received 26 May 2021; Received in revised form 23 November 2021; Accepted 9 June 2022
Available online 18 June 2022

dependent (STD) parameters have been suggested as a tool to describe the effects of intrinsic uncertainty on model states and output, and to possibly identify structural model deficits (Reichert and Mieleitner, 2009; Reichert et al., 2021). Indeed, by representing variability within the model structure, they appear particularly suited to account for our uncertain knowledge of the system at hand.

A clear benefit of modeling uncertainty with STD parameters is the consequential natural propagation of it through the model structure to the output. In fact, autocorrelation effects can be induced by the stochastic processes, and are also directed and amplified by the propagation of stochasticity throughout the model. Additionally, making process rate parameters stochastic is consistent with the fulfillment of mass balance equations, which might not be granted with other specific modeling strategies. Indeed, for example, if it is the level of the water in the reservoirs that is treated as a stochastic variable (Moradkhani et al., 2005; Vrugt et al., 2013), then mass balance equations are not exactly fulfilled. Importantly, STD parameters pose no theoretical limitations to the concurrent use of other complementary ways to treat uncertain knowledge. Observational error models on input and/or output, including models for systematic bias (Sikorska and Renard, 2017), as well as stochastic model states, can all be used concurrently with STD parameters.

Despite these appealing features, STD parameters are seldom applied to demanding modeling scenarios. This is mainly because of the intricacy of the algorithms needed to cope with the calibration of stochastic models, because of the computational resources required to achieve converged results, and possibly also because more complex and/or informative outcomes can be more challenging to mine and interpret. Hence, the overarching goal of this contribution is to comprehensively test and discuss the feasibility of using STD parameters for a non-linear multi-reservoir conceptual hydrological model applied to real-world data, and to show the ensuing implications on uncertainty quantification. By doing so, we advance previous studies by considering a more complex model, and by using up to 3 STD parameters at once. We also resort to a novel parallel framework (Šukys and Bacci, 2021), which allows us to benefit from available high-performance computing (HPC) infrastructure and to use the Particle Filter (PF) method coupled with a Markov Chain Monte Carlo approach to Bayesian inference (PMCMC) (e.g., Doucet and Johansen, 2009; Andrieu et al., 2010; Fearnhead and Künsch, 2018; Van Leeuwen et al., 2019). Although not pursued here, this could in principle be used to extend our work to non-linear stochastic processes, differently from what is possible with other methods, such as conditional Ornstein–Uhlenbeck sampling (Buser, 2003; Tomassini et al., 2009; Reichert and Mieleitner, 2009; Reichert et al., 2021). On the other hand, the calibration of stochastic models with the PMCMC method is usually more computationally expensive, and hence also likely to require expertise in parallel programming, or at least familiarity with and access to HPC resources. Thus, an additional goal of this contribution is to report on the algorithmic and computational aspects implicit in this more universal and wider-ranging approach to stochastic modeling in hydrology.

Ultimately, the main motivation for the present and possible future endeavors of this type, comes from the fact that approaches like this one can allow the modeler to improve the characterization of uncertainty by improving the partition of the variability into the different sources, and can help identify and correct possible model deficits (Reichert et al., 2021). Thus, we also list among our objectives the discussion of these aspects for our case study. With this aim in mind, we consider and discuss in particular the possible "misuse" of stochastic parameters, which occurs when the posterior dynamics of the stochastic parameters systematically adjusts for model deficits. Such a misuse can only occur, however, when the data are known and used to inform the dynamical behavior of the model, as in calibration/data-assimilation procedures. In contrast, compensation of model deficits simply cannot take place during prediction as observations are not used to inform the response of the model. This inevitably reduces the predictive power of those models for which misuse of stochastic parameters occurred in calibration, and this can be detected and quantified with cross-validation. Further steps, such as the explicit modeling of input uncertainty and the consideration of data sets with higher, e.g. hourly, time resolution, are all challenges that we suggest should be tackled in forthcoming studies.

In summary, the aims of this study are to explore, characterize, and discuss the application of STD parameters to a conceptual hydrological model structure by using the PMCMC method within a parallel computational framework. Efforts are specifically directed to determine the ensuing implications for uncertainty quantification and to assess possible modeling limitations of the proposed approach and of the underlying hydrological model, which the proposed approach can help to detect. In order to meet and communicate all the mentioned goals the following design is taken. First, we select and describe our hydrological case study including the observational data in Section 2.1, and the hydrological model that we use for its description in Section 2.2. Then we go into more details regarding STD parameters, inference, and PF method, see Sections 2.3,2.4,2.5,2.6. We close the methodological part with an overview on prior and post-processing, which hinges on the need to quantify and compare the results of the simulated models, both for calibration and cross-validation, Sections 2.7 and 2.8. Results are presented and interpreted in Section 3, and further discussed in 4. The contribution ends with laying down our Conclusions in Section 5.

2. Materials and methods

2.1. Study area and data

We apply our approach for inference and cross-validation to the Murg catchment, a small (80 km^2) pre-alpine foothill catchment in northeastern Switzerland, which is sometimes addressed to as Wängi catchment from the name of the village where the gauging station is located. This catchment belongs to the larger Thur river basin, and has been object of several previous studies (e.g., Dal Molin et al., 2020; Ammann et al., 2019; Schirmer et al., 2014), to which we address the reader for further specific hydro-geological information. Here, it suffices to say that the Murg area is characterized by steep slopes that can make the streamflow peaks quite sharp by draining directly into the river (Ammann et al., 2019), and that only 5% of precipitation is falling as snow, which implies that snowfall is not a prominent influencing factor of the streamflow (Dal Molin et al., 2020). Additionally, there are no natural or artificial reservoirs in the study area. All these notions are used to inform the building up of the hydrological model in Section 2.2.

The data used in this study consist of daily precipitation, potential evaporation, and streamflow for a 6 years period ranging from 01–09-1993 to 31–12-1999. As in previous works, (Dal Molin et al., 2020; Ammann et al., 2019), streamflow values come from the Swiss Federal Office for the Environment (FOEN), while precipitation and potential evapotranspiration from MeteoSwiss (2018). All data are pre-processed as described by Dal Molin et al. (2020). The selection of a daily temporal resolution is justified by a previous study (Dal Molin et al., 2020), and by the necessity to keep the computational time and budget within reasonable limits while still exploring multiple hydrological hypotheses. We use the observations from 01–09-1993 to 31–08-1997 for calibration, with a warm-up period to let the initial condition equilibrate equal to 365 days. The data from 01–09-1997 to 31–12-1999 serve for cross-validation.

2.2. Hydrological model and experiments

The lumped model structure used in this study is schematically shown in Fig. 1, and in terms of complexity and processes representation, is broadly reflective of typical conceptual rainfall-runoff models such as HBV (Lindström et al., 1997), GR4J (Perrin et al., 2003) or HyMod (Wagener et al., 2001).

More specifically, the model follows the classic three elements setup

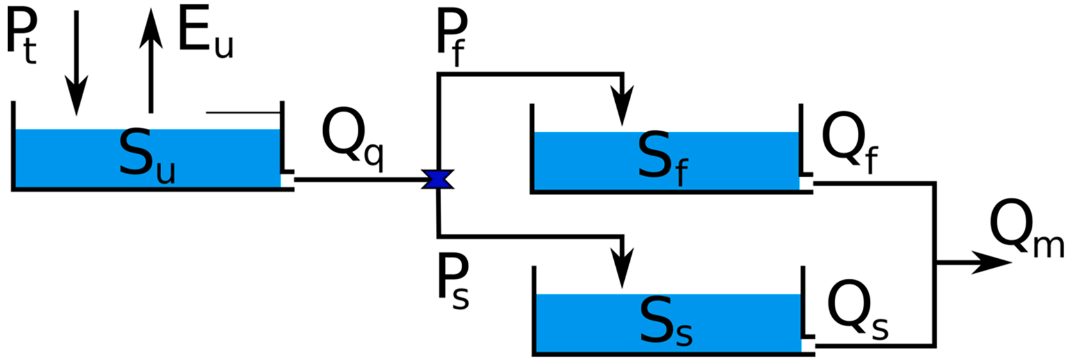


Fig. 1. Schematics of the conceptual hydrological model. The upstream unsaturated reservoir exchanges with the environment through precipitation P_t and evapotranspiration E_u terms. Its water level is S_u and its outflow feeds two downstream linear reservoirs through a splitting component. The final discharge Q_m is the sum of the outflow of the two downstream compartments.

outlined by Jakeman and Hornberger (1993), composed by an upstream element connected in series to two downstream units working in parallel. The upstream element is an "unsaturated" reservoir (UR, u subscript), and it exchanges directly with the environment through precipitation and evapotranspiration terms. The balance between these two determines the amount of water stored in UR, which controls, through a nonlinear term and a splitting unit, the flow to the two downstream elements. These last two units are both linear reservoirs, a "fast" reservoir (FR, f subscript) intended to represent the hydrograph peaks, and a "slow" reservoir (SR, s subscript) intended to represent baseflow. The model structure is built by using the SUPERFLEX modeling framework (Fenicia et al., 2011). We do not consider a lag function as in other models such as HBV as deemed unnecessary based on preliminary simulations. Supported by precipitation data, see Section 2.1, we also do not implement a snow component.

In order to limit the number of model calibration parameters, similarly to what done in some other conceptual models such as GR4J, some model parameters are fixed. In particular, we use β=5 unless otherwise stated, which approximates a smooth threshold behavior for discharge generation and is justified by preliminary simulations, see Section 3.1. We also use $m = 0.01$ as in previous works (Fenicia et al., 2013; Fenicia et al., 2018), which implies that actual evaporation is approximately equal to the potential evaporation unless the reservoir is close to empty.

The dynamics of a model with structure as in Fig. 1 is controlled by Eqs. (1):

$$\begin{cases} \dfrac{dS_u}{dt} = P_t - Q_q - E_u = P_t - P_t\left(\dfrac{S_u}{S_{uMax}}\right)^{\beta} - C_e E_p \dfrac{S_u/S_{uMax}(1+m)}{S_u/S_{uMax}+m} \\ \dfrac{dS_f}{dt} = P_f - Q_f = \left(1-D\right)Q_q - k_f S_f = \left(1-D\right)P_t\left(\dfrac{S_u}{S_{uMax}}\right)^{\beta} - k_f S_f \\ \dfrac{dS_s}{dt} = P_s - Q_s = DQ_q - k_s S_s = DP_t\left(\dfrac{S_u}{S_{uMax}}\right)^{\beta} - k_s S_s \\ Q_m = Q_f + Q_s = k_f S_f + k_s S_s \end{cases} \tag{1}$$

From eq. (1), S_{uMax} controls the maximum water level in UR (anything above that level is spilled out to the downstream units in a single time step as in a truly on–off model), C_e is the scaling coefficient for the evapotranspiration term, D is the split parameter that controls how the water from UR is partitioned between the two downstream reservoirs (the larger D the larger the flux to the slow reservoir), and k_f and k_s are the coefficients that control the linear storage-discharge relationship in the corresponding reservoir (FR and SR, respectively).

In the following case studies, parameters k_f, k_s, and/or D can be either deterministic (meaning, constant in time) or stochastic (meaning, represented by a suited time-dependent stochastic process, see Section 2.3). In particular, we investigate a fully deterministic model and 4 stochastic models with different selections of STD parameters. The 5 experiments and the associated model calibration parameters are schematically shown in Table 1. The deterministic model *Det* does not include any stochastic process, hence it only numbers 5 fitted parameters that we summarize in vector $\theta^m = (S_{uMax}, C_e, D, k_f, k_s)$. In model *Sto* $-K_f$ the STD parameter is k_f, in *Sto* $-K_s$ it is k_s and in *Sto* $-D$ is D. Finally, in *Sto* $-KKD$ we make all of them concurrently stochastic. In the stochastic models, the temporal dynamics of the STD parameters is inferred jointly with both the parameters of the processes that model them and with the other model parameters.

The choice of making k_f, k_s, and/or D stochastic rests on both theoretical and practical aspects. Indeed, by limiting stochastic modeling to process rates, mass balance equations formulated in the deterministic model remain valid. This would be more difficult for capacity parameters such as maximum reservoir levels. The three parameters mentioned above are the key process rate parameters.

2.3. STD parameters

By denoting a general STD parameter with θ_s, we model the evolution of a suited transformation of that parameter $f(\theta_s)$, with a linear time-continuous autoregressive stochastic process $\chi^{OU}_{f(\theta_s)}$ called Ornstein–Uhlenbeck (OU) process (Uhlenbeck and Ornstein, 1930). This formally writes as $f(\theta_s(t)) = \chi^{OU}_{f(\theta_s)}(t)\ \forall t$. Practically, this means that, given the value of the transformed STD parameter θ_s at time t_0, $f(\theta_s(t_0))$, then the probability distribution of its subsequent value $f(\theta_s(t_1))$ at time $t_1 > t_0$ is given by the Normal distribution:

$$f\left(\theta_s\left(t_1\right)\right)|f\left(\theta_s\left(t_0\right)\right) \sim N\left(\mu^{OU}_{f(\theta_s)} + \left(f\left(\theta_s\left(t_0\right)\right) - \mu^{OU}_{f(\theta_s)}\right)\exp\left(-\frac{t_1-t_0}{\tau^{OU}_{f(\theta_s)}}\right), \sigma^{OU}_{f(\theta_s)}\sqrt{1-\exp\left(-2\frac{t_1-t_0}{\tau^{OU}_{f(\theta_s)}}\right)}\right) \tag{2}$$

From eq. (2), the parameters of the OU process are its asymptotic mean $\mu^{OU}_{f(\theta_s)}$, its asymptotic standard deviation $\sigma^{OU}_{f(\theta_s)}$, and its autocorrelation time $\tau^{OU}_{f(\theta_s)}$. We summarize these parameters in the parameter vector $\theta^{OU} = \left(\mu^{OU}_{f(\theta_s)}, \sigma^{OU}_{f(\theta_s)}, \tau^{OU}_{f(\theta_s)}\right)$ and we keep the notation θ^{OU} also for the cases with multiple stochastic parameters where this parameter vector contains the means, standard deviations and correlation times of all the STD parameters.

While an OU process has unbounded codomain, model parameters might not exceed a given range of values. This is the reason why we generally need the transformation f on θ_s. Such a transformation allows us to map the usually limited support of the modeled parameter onto the values that the OU process can take, $\chi^{OU}_{f(\theta_s)}(t) \in \mathbb{R}$. In this study, we make model parameters k_f,k_s, and D stochastic. Since k_f and k_s are constrained to $\mathbb{R}^+$, we need a transformation $f(k_f) : \mathbb{R}^+ \rightarrow \mathbb{R}$. To this aim, we simply choose the natural logarithm as in previous works (Reichert et al., 2021; Reichert and Mieleitner, 2009). This means that what we describe with the OU process is not k_f or k_s, but it is their natural logarithms, e.g., $\ln(k_f(t)) = \chi^{OU}_{\ln(k_f)}(t)$. The equation for k_s is analogous. For the split parameter, the transformation that we choose is the logit function. This is because $D \in [0,1]$, and the logit function is a possible way to map this interval to $\mathbb{R}$: $\text{logit}\left(D\left(t\right)\right) = \ln\left(\frac{D(t)}{1-D(t)}\right) = \chi^{OU}_{\text{logit}(D)}\left(t\right)$.

In summary, when a parameter is modeled by an OU process, we have to infer the value of the asymptotic mean of the process $\mu^{OU}_{f(\theta_s)}$, its asymptotic standard deviation $\sigma^{OU}_{f(\theta_s)}$, and its autocorrelation time $\tau^{OU}_{f(\theta_s)}$, jointly with the posterior of its actual time course. To simplify the notation and to highlight the fact that we always refer to the back-transformed parameters θ_s, we use the symbol $< \theta_s >$ to indicate the time-mean of $f^{-1}(\chi^{OU}_{f(\theta_s)}(t))$where appropriate in Section 3. We now proceed by considering the error model that we use, to then be able to discuss our inference framework.

2.4. Observational and lumped error models

In our setup, the output error term applies to the discharge of the hydrological model in the same way regardless of the type of process model (deterministic or stochastic). The parameterization that we choose is also the same. These choices are supported by simplicity, by the willingness to be able to perform direct comparisons, and by the flexibility of the selected error parameterization. However, we should notice that the error term on the output has in principle a different meaning for the deterministic or stochastic models. In the first case, the error lumps together all sources of uncertainty, while for stochastic models it ideally just represents the observation error, as parametric, model structure uncertainty, and intrinsic stochasticity are considered by the STD parameters.

In the remainder of the text, for simplicity, we call the output error model observational likelihood even when, for the deterministic hydrological model, it represents the distribution of observations resulting from intrinsic uncertainty in addition to observation error.

To model the lumped or observational error for the streamflow Q_{obs}, we opt for a widespread approach. As it is well known, the precision of the measurement depends on the magnitude of the streamflow itself. A possible way to deal with such a heteroscedastic error term is to transform the data and the model output via a Box-Cox (BC) transformation (Box and Cox, 1964), and to then apply a homoscedastic Gaussian error model in the transformed space (McInerney et al., 2017):

$$\begin{aligned} &BC(Q_{obs}) = BC(Q_m) + \epsilon_{BC} \\ &\epsilon_{BC} \sim N(0, \sigma_{BC}) \\ &\rightarrow BC(Q_{obs}) \sim N(BC(Q_m), \sigma_{BC}) \end{aligned} \tag{3}$$

where $BC\left(Q\right) = \frac{Q^{\lambda}-1}{\lambda}$, Q_m is the streamflow output by the hydrological model, Q_{obs} is the observed discharge, and the last expression in (3) provides the expected distribution of the observed data given the model output. In other words, it provides the basic and actionable definition of the observational likelihood given data and model output, definition that is required to calibrate the deterministic model in a Bayesian setting, see Section 2.5. We note that the observational error model (3) adds 2 parameters per hydrological model to the ones in Table 1: the standard deviation of the normal distribution in BC space, σ_{BC}, and the parameter λ of the transformation. As it is commonplace, we inferr σ_{BC} while fixing λ to 0.5 as recommended in previous work (McInerney et al., 2017). Our parameter vector for the error model thus consists of a single element, $\theta^y = (\sigma_{BC})$. We also note that we use the back-transformed error $\epsilon = \text{Q}_{\text{obs}} - Q_m = BC^{-1}(BC(Q_m) + \epsilon_{BC}) - Q_m$ to expose the relevant results in Section 3.

In order not to also have to resolve identifiability problems between input and intrinsic uncertainties, we do not consider input uncertainty explicitly, but analyse potential implicit effects of input uncertainty in our results. We also note that input uncertainty could be considered in the used framework, for example by modelling precipitation as a transformed stochastic process and using the observations to condition it (Del Giudice et al., 2016).

All our models, namely the hydrological model, see Eqs. (1) and also Table 1, and the observational or lumped error model, eq. (3), have parameters, which are grouped in the vectors θ^m and θ^y, respectively. These parameters are inferred based on the available data, as detailed in the following Section.

2.5. Inference

The parameters of the deterministic hydrological model given by Eq. (1) and the lumped error model given by Eq. (3) are inferred from the data using Bayes equation:

$$P\left(\theta|Q_{obs}\right) = \frac{P(Q_{obs}, \theta)}{P(Q_{obs})} = \frac{L(Q_{obs}|\theta)\pi(\theta)}{E(Q_{obs})} \propto L\left(Q_{obs}|\theta\right)\pi\left(\theta\right). \tag{4}$$

Here, $\theta = (\theta^m, \theta^y)$ are the parameters of both the hydrological and error models, the joint probability density of data and parameters $P(Q_{obs}, \theta)$ is made explicit as the customary product between likelihood $L(Q_{obs}|\theta)$ and prior $\pi(\theta)$, and the the unconditional probability density of the observations $P(Q_{obs})$ is simply renamed evidence $E(Q_{obs})$ as it is commonplace. The likelihood function of the model $L(Q_{obs}|\theta)$ is the probability density of the observed data Q_{obs} given the parameters θ, and used as a function of the parameters by substituting the actual observations for the corresponding argument, $\pi(\theta)$ summarizes previous knowledge about plausible parameters values, and the evidence results from the joint distribution through marginalization over the parameters, $E(Q_{obs}) = \int L(Q_{obs}|\theta)\pi(\theta)\text{d}\theta$, hence it is constant for a given model.

Although the prior is an important component of Bayesian inference, see Section 2.7, the core prior element are the modeling choices reflected by the likelihood. The error model in eq. (3), which assumes independence between model residuals, implies the following likelihood:

$$L\left(Q_{obs}|\theta\right) = L_{t_1:t_n}\left(Q^{t_1:t_n}_{obs}|\theta\right) = \prod_{k=1}^{n} p^Q_{t_k}\left(Q^{t_k}_{obs}|Q^{t_k}_m(\theta^m), \theta^y\right) \tag{5}$$

where

Table 1
Parameters present in the models. "–" means that the parameter is absent in the relevant model. Parameters β and m in Eqs. (1) are kept fixed in all models to 5 and 0.01 respectively, unless stated otherwise.

	S_{uMax}	C_e	k_f	k_s	D	$\chi^{OU}_{f(k_f)}$	$\mu^{OU}_{f(k_f)}$	$\sigma^{OU}_{f(k_f)}$	$\tau^{OU}_{f(k_f)}$	$\chi^{OU}_{f(k_s)}$	$\mu^{OU}_{f(k_s)}$	$\sigma^{OU}_{f(k_s)}$	$\tau^{OU}_{f(k_s)}$	$\chi^{OU}_{f(D)}$	$\mu^{OU}_{f(D)}$	$\sigma^{OU}_{f(D)}$	$\tau^{OU}_{f(D)}$
Det	✓	✓	✓	✓	✓	–	–	–	–	–	–	–	–	–	–	–	–
Sto-K_f	✓	✓	✓	✓	✓	✓	✓	✓	✓	–	–	–	–	–	–	–	–
Sto-K_s	✓	✓	✓	✓	✓	–	–	–	–	✓	✓	✓	✓	–	–	–	–
Sto-D	✓	✓	✓	✓	✓	–	–	–	–	–	–	–	–	✓	✓	✓	✓
Sto-KKD	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

$$\underset{t_k}{\overset{Q}{p}}\left(Q_{obs}^{t_k}\middle|Q_m^{t_k}(\theta^m),\theta^y\right) = \rho_{N\left(BC\left(Q_m^{t_k}(\theta^m)\right),\sigma_{BC}\right)}\left(BC\left(\underset{obs}{\overset{t_k}{Q}}\right)\right)\cdot\frac{dBC(Q)}{dQ}\Bigg|_{Q=Q_{obs}^{t_k}} = \rho_{N\left(BC\left(\underset{m}{\overset{t_k}{Q}}\left(\theta^m\right)\right),\ \sigma_{BC}\right)}\left(BC\left(\underset{obs}{\overset{t_k}{Q}}\right)\right)\cdot\underset{obs}{\overset{t_k}{Q}}\lambda-1 \tag{6}$$

Here $p_{t_k}^Q(Q_{obs}^{t_k}\Big|Q_m^{t_k}(\theta^m),\theta^y)$ is the probability that we measure $Q_{obs}^{t_k}$ at time t_k given the output of the model $Q_m^{t_k}$ at that time, with $\rho_{N(\mu,\sigma)}(x)$ the probability density of a Normal distribution with mean μ and standard deviation σ evaluated at x. This expression is substituted into Eq. (4) to complete our definition of the posterior for the deterministic model.

If one or multiple parameters of the model are chosen to be stochastic, however, we get a hierarchical model. In this case, we denote the stochastic parameter(s) (a subset of the parameters θ^m of the hydrological model) by θ_s^m, and the remaining constant parameters by $\theta_{-s} = (\theta_{-s}^m, \theta^y, \theta^{OU})$, where θ_{-s}^m are the constant parameters of the hydrological model, θ^y are the parameters of the error model, and $\theta^{OU} = \left(\mu_{f(\theta_s)}^{OU}, \sigma_{f(\theta_s)}^{OU}, \tau_{f(\theta_s)}^{OU}\right)$ are the parameters of the Ornstein–Uhlenbeck process(es), see Section 2.3. We then have to modify Eq. (4) to:

$$P\left(\theta_{-s},\theta_s^m|Q_{obs}\right) = \frac{L_{t_1:t_n}\left(Q_{obs}^{t_1:t_n}|\theta_{-s},\theta_s^m\right)p\left(\theta_s^m|\theta^{OU}\right)\pi(\theta_{-s})}{E(Q_{obs})} \propto L_{t_1:t_n}\left(Q_{obs}^{t_1:t_n}|\theta_{-s},\theta_s^m\right)p\left(\theta_s^m|\theta^{OU}\right)\pi\left(\theta_{-s}\right). \tag{7}$$

Here, $p(\theta_s^m|\theta_s^{OU})$ is the back-transformed joint probability density of the Ornstein–Uhlenbeck process at all time points of a grid with fine resolution. For a single stochastic parameter, this joint density is given as the product of an unconditional normal distribution for the initial point multiplied by the product of the conditional distributions of the next point given the previous one, see eq. (2), and multiplied by the correction factor for back-transformation:

$$p\left(\theta_s^m|\theta^{OU}\right) = \rho_{N\left(\mu_{f(\theta_s^m)}^{OU},\sigma_{f(\theta_s^m)}^{OU}\right)}\left(f\left(\theta_s^m(t_1)\right)\right)\prod_{k=2}^{n}\rho_{N(\mu_k,\sigma_k)}\left(f\left(\theta_s^m(t_k)\right)\right)\cdot\prod_{k=1}^{n}\frac{df(\theta)}{d\theta}\Big|_{\theta_s^m(t_k)}$$

with

$$\mu_k = \mu_{f(\theta_s^m)}^{OU} + \left(f\left(\theta_s^m\left(t_{k-1}\right)\right) - \mu_{f(\theta_s)}^{OU}\right)\exp\left(-\frac{t_k-t_{k-1}}{\tau_{f(\theta_s^m)}^{OU}}\right), \quad \sigma_k = \sigma_{f(\theta_s^m)}^{OU}\sqrt{1-\exp\left(-2\frac{t_k-t_{k-1}}{\tau_{f(\theta_s^m)}^{OU}}\right)} \tag{8}$$

Note that, for multiple stochastic parameters, we would need a product of expressions of the form given by Eq. (8). Note also that we use the same notation for the time steps in Eq. (8) as in the observation likelihood (5), since this is the implementation used in this paper. However, this is not a requirement of this technique. The time discretization of the stochastic process is required only to integrate the differential equations of the hydrological model, hence it is completely independent of the observation time spacing. In Eq. (7), $\pi(\theta_{-s})$ is the joint prior of the constant parameters and will usually be assumed to be the product of independent distributions of the three categories of parameters, $\pi(\theta_{-s}) = \pi(\theta_{-s}^m)\pi(\theta^y)\pi(\theta_s^{OU})$ (independence assumptions are often taken also between the parameters within these three categories), and $E(Q_{obs})$ is now given by $\iint L(Q_{obs}|\theta_{-s},\theta_s^m)p(\theta_s^m|\theta_s^{OU})\pi(\theta_{-s})\mathrm{d}\theta_s^m\mathrm{d}\theta_{-s}$.

Since the parameterization of the error model is the same for all process models, we note that (5) and (6) (with $(\theta_{-s}^m, \theta_s^m)$ for θ^m and $(\theta_{-s}, \theta_s^m)$ for θ) are still valid for $L_{t_1:t_n}(Q_{obs}^{t_1:t_n}|\theta_{-s}, \theta_s^m)$ in (7), once Q_m is available at each time step t_k by simple time-integration of the hydrological model and of the respective stochastic process(es), the propagation of which in time evaluates (8).

From (7) we get the marginal posterior of the constant parameters by integrating out the stochastic dynamics θ_s^m:

$$P\left(\theta_{-s}|Q_{obs}^{t_1:t_n}\right) \propto \int L_{t_1:t_n}\left(Q_{obs}^{t_1:t_n}|\theta_{-s},\theta_s^m\right)p\left(\theta_s^m|\theta_s^{OU}\right)d\theta_s^m\cdot\pi\left(\theta_{-s}\right) = L_{t_1:t_n}^{marg}\left(Q_{obs}^{t_1:t_n}|\theta_{-s}\right)\pi\left(\theta_{-s}\right) \tag{9}$$

with marginal likelihood:

$$L_{t_1:t_n}^{marg}\left(Q_{obs}^{t_1:t_n}|\theta_{-s}\right) = \int L_{t_1:t_n}\left(Q_{obs}^{t_1:t_n}|\theta_{-s},\theta_s^m\right)p\left(\theta_s^m|\theta_s^{OU}\right)d\theta_s^m. \tag{10}$$

Note that the last expression in Eq. (9) is again of the form of (4). However, the (marginal) likelihood is now the very high-dimensional integral (10), which implies the need to infer the time series of θ_s^m. This makes inference for a stochastic model a much harder problem than inference for a deterministic one.

2.6. Numerical approach

For the deterministic model we use a Markov Chain Monte Carlo (MCMC) approach to sample from the posterior given by Eq. (4) with the likelihood defined by Eqs. (5) and (6), and choose the affine invariant Markov chain Monte Carlo ensemble (EMCEE) sampler by Foreman-Mackey et al. (2013) to achieve fast convergence.

When using stochastic parameter(s) we apply the Particle Markov Chain Monte Carlo (PMCMC) scheme described in Andrieu et al. (2010). This approach consists of combining a Particle Filter (PF) (e.g., Doucet and Johansen, 2012; Fearnhead and Künsch, 2018; van Leeuwen et al., 2019) to sample the posterior marginals of the stochastic parameter(s) at each time point conditional on the constant parameters (and on past to present observations), and using these conditional samples to approximate the marginal likelihood (10). The approximate marginal likelihood is then used within an MCMC scheme to sample the constant parameters, including the parameters of the stochastic process, according to Eq. (9). Again, for this last step, we use the EMCEE sampler cited above. We provide some more details on this PMCMC approach in the next paragraph.

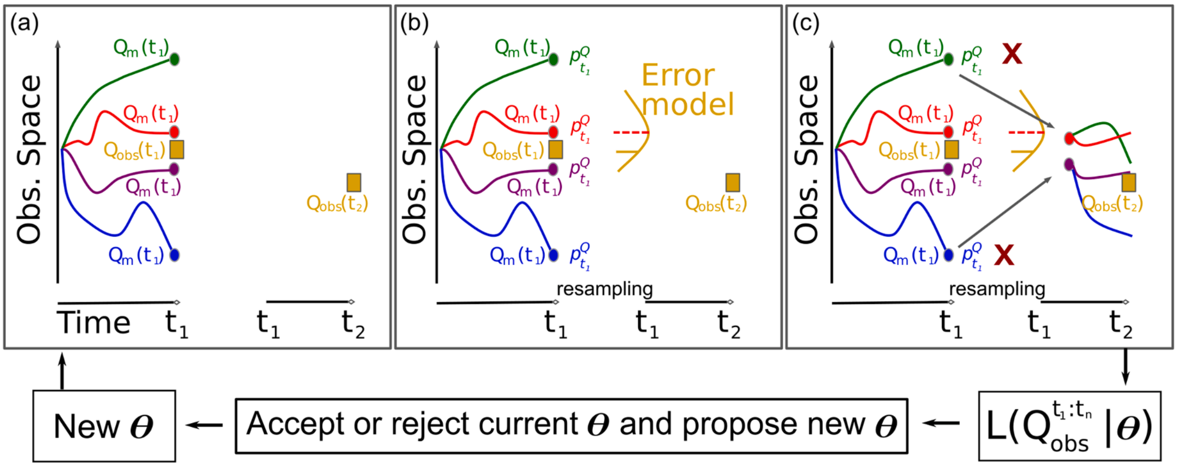

(a)
Obs. Space
$Q_m(t_1)$
$Q_{obs}(t_1)$
$Q_{obs}(t_2)$
Time
t_1
t_1
t_2
(b)
Error model
$p^Q_{t_1}$
resampling
(c)
X
New $\boldsymbol{\theta}$
Accept or reject current $\boldsymbol{\theta}$ and propose new $\boldsymbol{\theta}$
$L(Q^{t_1:t_n}_{obs}|\boldsymbol{\theta})$

Fig. 2. Sketch of the coupling between a Metropolis MCMC scheme with an inner Particle Filter. At each observational time, the (instantaneous) observational likelihood $p^Q_{t_k}$ of each model execution (particle) is used to fully resample the particles ensemble. In this sketch, the green and blue particles are far from the data, hence their likelihood value is small, hence they are not resampled. The imagined result is that they are restarted by cloning the purple and red particles, which are envisaged to be the resampled particles. Propagation then proceeds independently to the next observation point, and trajectories diverge due to the stochasticity of the model. At the end the (marginal) likelihood is used to evaluate whether or not to reject the current proposed parameters θ.

Fig. 2 provides an overview of the PMCMC scheme to the joint inference of constant and stochastic parameters.

PMCMC (Andrieu et al., 2010) starts with sampling posterior marginal states of the stochastic parameter θ^m_s at each time point for given constant parameters by the PF. As illustrated in Fig. 2, this is done by sampling values $\theta^m_s(t_i)$ at a given time point from samples of points (particles) at the previous time point, t_{i-1}, integrating the hydrological model for each of these trajectories within the interval from time t_{i-1} to time t_i, and then calculating importance weights proportional to the observation likelihood at the new time point. The values (particles) are then resampled according to the weights (observational likelihoods), and the weights are used to calculate the approximate marginal likelihood. To derive this approximate marginal likelihood, we first write the marginal likelihood (10) as a product of an unconditional distribution for the first observed discharge multiplied by conditional distributions for discharge at subsequent time points:

$$L^{marg}_{t_1:t_n}\left(Q^{t_1:t_n}_{obs}|\theta_{-s}\right) = p\left(Q^{t_1}_{obs}|\theta_{-s}\right)\prod_{k=2}^{n} p\left(Q^{t_k}_{obs}|Q^{t_1:t_{k-1}}_{obs},\theta_{-s}\right), \qquad (11)$$

where n is the number of observations. This equation allows us to calculate approximate marginal likelihoods, $\widehat{L}$, based on samples $\{\theta^m_{s,i}(t_k)\}^{N_p}_{i=1}$ of the marginal distributions at all time points from the Particle Filter:

$$\widehat{L}^{marg}_{t_1:t_n}\left(Q^{t_1:t_n}_{obs}|\theta_{-s}\right) = \prod_{k=1}^{n}\frac{1}{N_p}\sum_{i=1}^{N_p} p^Q_{t_k}\left(Q^{t_k}_{obs}\middle|Q^{t_k}_{m}\left(\theta^m_{-s},\theta^m_{s,i}\left(t_1:t_k\right)\right),\theta^y\right) \qquad (12)$$

where N_p is the number of particles and $p^Q_{t_k}$ is given by (6) in this work. There are two main reasons why this expression is approximate. First, the samples of the time series of θ^m_s are replaced by the samples from the marginals at each time point that are only conditioned on the data up to this time point (this is a property of the filtering approach) and second, with finite samples, distributional properties can only be calculated approximately. In an "outer loop", the approximation given by Eq. (12) can be used in (9) to sample from the marginal posterior of the constant parameters, θ_{-s}, using the EMCEE sampler.

All calculations are carried out using the SPUX framework by Šukys and Bacci (2021). In each simulation, if not differently stated, we use 24

Table 2
Prior distributions for the model of the physical system and for the observational error model of the output. LN(a,b) stands for log-normal distribution with mean a and coefficient of variation b. N(a,b) stands for normal distribution with mean a and standard deviation b. α in $\sigma^{OU}_{logit(D)}$ has been numerically estimated by using the logit transformation over a large number of samples drawn from the prior of D of the deterministic model to estimate the variance of D in OU space ($\alpha = 1.636077$).

	Det	*Sto* $-K_F$	*Sto* $-K_S$	*Sto* $-D$	*Sto* $-KKD$
S_{uMax}	LN(200,1)	LN(200,1)	LN(200,1)	LN(200,1)	LN(200,1)
C_e	LN(1,1)	LN(1,1)	LN(1,1)	LN(1,1)	LN(1,1)
k_f	LN(5,1)	–	LN(5,1)	LN(5,1)	–
k_s	LN(10^{-3},1)	LN(10^{-3},1)	–	LN(10^{-3},1)	–
D	LN(0.5,0.5)	LN(0.5,0.5)	LN(0.5,0.5)	–	–
$\mu^{OU}_{\ln(k_f)}$	–	$\mathrm{N}\left(\ln(5)-0.5\ln(2),\sqrt{\ln(2)}\right)$	–	–	Infer
$\sigma^{OU}_{\ln(k_f)}$	–	LN(1,1)	–	–	LN(1,1)
$\tau^{OU}_{\ln(k_f)}$	–	LN(12,1)	–	–	LN(12,1)
$\mu^{OU}_{\ln(k_s)}$	–	–	$\mathrm{N}\left(\ln(10^{-3})-0.5\ln(2),\sqrt{\ln(2)}\right)$	–	$\mathrm{N}\left(\ln(10^{-3})-0.5\ln(2),\sqrt{\ln(2)}\right)$
$\sigma^{OU}_{\ln(k_s)}$	–	–	LN(1,1)	–	LN(1,1)
$\tau^{OU}_{\ln(k_s)}$	–	–	LN(12,1)	–	LN(12,1)
$\mu^{OU}_{logit(D)}$	–	–	–	N(0,1)	N(0,1)
$\sigma^{OU}_{logit(D)}$	–	–	–	LN(α,1)	LN(α,1)
$\tau^{OU}_{logit(D)}$	–	–	–	LN(12,1)	LN(12,1)
σ_{BC}	LN(1,1)	LN(1,1)	LN(1,1)	LN(1,1)	LN(1,1)

particles and 40 Markov chains propagated in parallel as a trade-off between accuracy and computational load and time.

2.7. Prior

Table 2 collects the prior marginal distributions for all the parameters that we infer. The joint prior is constructed by assuming independence. As it is often the case, with the exception of parameter D and of the parameters that characterize the OU process and the error model, we sample in log space. This means in practice that prior log-normal distributions are transformed into the corresponding normal distributions in the numerical implementation. To start the Markov chains, we draw from Gaussian distributions centered at the mean of the marginal priors and with standard deviation equal to 10% the value of the mean. As a general note, we use quite lenient priors in terms of support, albeit with clear preferences for those specific values that expert knowledge suggests. Regarding the initial states, as mentioned in Section 2.1, we use a 1 year long warm-up phase at the beginning of each model execution to allow the level of the water in the initially empty reservoirs to adapt.

2.8. Analysis of Results

We distinguish three main steps for the analysis of results:

1. Analysis of convergence of the Markov chains, Section 3.1.
2. Model deficit analysis, Section 3.2.
3. Posterior analysis and prediction, Section 3.3.

The first step is a prerequisite for any interpretation of the results. In case of poor convergence, we would have to improve the sampling procedure or extend the Markov chains. The second step is intended to establish whether the additional degrees of freedom of the stochastic parameters are misused during calibration to compensate for model deficits. If this is the case, the statistical properties of the inferred parameter time series are not valid and either the model would have to be improved or the parameters leading to these problems would have to be kept constant. We check this by cross-validation and by analyzing the inferred parameter time-series. Finally, after these checks have been passed, we can move on to the third step of interpretation of the results and prediction. We discuss the methodologies applied in these three steps in the following three subsections.

2.8.1. Convergence analysis

In most cases, visual analysis of the Markov chains provide sufficient information of convergence deficits due to extended burn-in periods, weak mixing, or long residence time in secondary modes. Visual inspection also provides insights that can be useful in resolving the convergence issue. On the other hand, quantitative convergence measures, such as the estimation of the effective sample size and standard MCMC convergence tests, can be very useful for getting indications on which chains need a more detailed inspection. Although we make use of both approaches in Section 3.1, the most satisfactory corroboration of convergence comes from diagnostic runs, where we change the models slightly, and achieve marginal posteriors that are very similar to the original ones. We obtain this while testing the influence of the number of particles on model $Sto-K_f$, Section 3.1, and of the prior for $\tau^{OU}_{\ln(k_s)}$ on model $Sto-K_s$, Section 3.2.3.

2.8.2. Model deficit analysis

The strongest methodology to identify misuse of the degrees of stochasticity to compensate for model deficits is cross-validation, which means that the model is run with the calibrated parameters without assimilating the available output data, which are used exclusively to asses the model's predictive power. This check is particularly important for stochastic models, where during calibration the time course of the time dependent parameters is inferred. In such a case, the difference in model performance between calibration and cross-validation can be indicative of the extent to which the time dependent parameters compensate for model deficits. If during inference the time-dependence of a stochastic parameter is used to compensate for systematic model deficits, a good performance will not be reproducible in cross-validation due to the (uninformed) random evolution of the stochastic process. For this reason, predictions will be poor and cross-validation will make the problem, if present, identifiable.

As shown in Section 3.2.1, aside from visual inspections, to assess cross-validation results we use four summary metrics, similarly to previous work (Reichert et al., 2021). Those are the relative spread, σ_{rel}, the distribution of the values that Q_{obs} attains within the cumulative distribution function relevant to the model outputs Q_m, which we simply refer to as CDF, the well-known Nash–Sutcliffe efficiency (NSE) (Nash and Sutcliffe, 1970), and the flashiness index (FI) (Baker et al., 2004; Fenicia et al., 2018).

The relative spread is a diagnostic metric that aims to inform on the width of the distribution of the model output Q_m. It does so by taking the average of the ratio between the standard deviation (SD) of the model predictions at a given time and the value of the corresponding observation. For a stochastic model, we can in principle partition the data by grouping according to particles, parameters, or both. Namely, at any observational time t_i, we can compute the SD across parameters, keeping the particles separated, across particles keeping the parameters separated, Eq. (13b), or across all the data, Eq. (13a):

$$\sigma_{rel}\left(Q_m, Q_{obs}\right) = \begin{cases} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \dfrac{SD[Q_m(t_i)]_{\text{parameters,particles}}}{Q_{obs}(t_i)} & (a) \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \dfrac{SD[Q_m(t_i)]_{\text{particles}}}{Q_{obs}(t_i)} & (b). \end{cases} \tag{13}$$

Note that with Eq. (13a) we obtain one single scalar, while with (13b) we get as many σ_{rel} values as there are parameters samples. Hence, we can display a distribution of values. In general, we find that grouping the data by the parameters does not add much to the description obtained by Eq. (13a), and so we do not include those results for brevity.

We examine the CDF, eq. (14), to complement and extend the analysis by σ_{rel}, as this is another metric that reports on the match and span of the simulation data with respect to the observations. Indeed, if the observations are drawn from the model realizations, then the respective CDF values should delineate an uniform distribution. Any deviation hints to specific issues. An excessive weight on any of the tail would point to coverage problems, as such a shape would manifest that observations are too often outside the range of model's predictions. In contrast, distributions peaked around the center would suggest that either overparameterization is at play (especially if the respective value of σ_{rel} is small), or that there might be an excess of uncertainty (especially if the corresponding σ_{rel} is large). For simplicity, for this metric, we just group all the data together, as the metric itself amounts to a distribution:

$$CDF\left(Q_m, Q_{obs}\right) = F_{[Q_m(t_i)]_{\text{parameters,particles}}}\left(Q_{obs}\left(t_i\right)\right). \tag{14}$$

The NSE is likely the most used metric in hydrology and is defined by Eq. (15a), where $E[\cdot]$ is the expected value operator:

$$NSE\left(Q_m, Q_{obs}\right) = \begin{cases} E\left[1 - \dfrac{\sum_{i=1}^{N_{obs}} \left(Q_m(t_i) - Q_{obs}(t_i)\right)^2}{\sum_{N_{obs}}^{i=1} \left(Q_{obs}(t_i) - E[Q_{obs}]\right)^2}\right] & (a) \\ 1 - \dfrac{\sum_{i=1}^{N_{obs}} \left(Q_m(t_i) - Q_{obs}(t_i)\right)^2}{\sum_{N_{obs}}^{i=1} \left(Q_{obs}(t_i) - E[Q_{obs}]\right)^2} & (b). \end{cases} \tag{15}$$

Essentially, the NSE is a measure of fit of the predictions to the data. Acceptable model performance customarily implies values $>= 0.5$. In principle, each model trajectory, each of which is generated by a specific combination of particles and parameters, has its own NSE value. In Section 3.2.1 we present the distribution for all those values, Eq. (15b), and their expected value (15a).

The FI provides a measure of the fluctuation of models predictions, Eq. (16a):

$$FI\left(Q_m\right) = \begin{cases} E\left[\dfrac{\dfrac{1}{N_{obs}-1}\displaystyle\sum_{i=2}^{N_{obs}}\left|Q_m\left(t_i\right) - Q_m\left(t_{i-1}\right)\right|}{\displaystyle\sum_{i=1}^{N_{obs}} Q_m\left(t_i\right)}\right] & (a) \\ \dfrac{\dfrac{1}{N_{obs}-1}\displaystyle\sum_{i=2}^{N_{obs}}\left|Q_m\left(t_i\right) - Q_m\left(t_{i-1}\right)\right|}{\displaystyle\sum_{i=1}^{N_{obs}} Q_m\left(t_i\right)} & (b). \end{cases} \quad (16)$$

Similarly to the NSE, each model trajectory has in principle its own scalar FI value, (16b), and this is what we plot in the relevant distributions in Section 3.2.1, together with the average value (16a).

For all the metrics, we never resort to smoothing of trajectories in post-processing. This means that the state of the model for a model's trajectory is always the one that emerges from the resampling scheme set by the PF up to the relevant observational time point, see Section S1 and Table S1 in the Supplementary Material. If the quantitative assessment of cross-validation results points to misuse of the STD parameter (s), to determine the cause of the problem, an analysis of the posterior time-series of the stochastic parameter(s) is very useful, see Section 3.2.2. This can lead to the identification of dependencies of the STD parameter(s) on external influence factors or internal model states, Section 3.2.2, and/or can help formulating additional hypotheses, Section 3.2.3. For our case study, to perform cross-validation tests, we draw 400 parameters samples from the joint parameters posterior distribution, select the associated model state at the end of inference, and continue the simulation for a couple of years time. For stochastic models, we still run as many executions per sample as we do for inference, *viz.*, we run 24 cross-validation simulations per parameter sample.

2.8.3. Posterior analysis and prediction

Once convergence is established and the stochastic model is either improved to avoid misuse of the STD parameters during calibration to compensate for model deficits, or those parameters are kept constant, or there are not such problems, the model can be used for conventional posterior analysis, such as assessing marginal posteriors, 3.3.1–3.3.2 and prediction time series, Section 3.3.3. Due to the explicit consideration of intrinsic uncertainty by the stochastic approach, we expect a more realistic description of the uncertainty in internal model states than for a lumped error model that only adds the uncertainty to the final outcome. We also expect a much more realistic distinction of (smaller) uncertainty during the calibration period when we condition the model to the observations, compared to (larger) uncertainty for prediction when we do not have observed information about the model state, although this limitation of the deterministic approach could be alleviated by resorting to a more complex formulation of the lumped error model (Reichert and Schuwirth, 2012).

2.8.4. Post-processing tools

All the routines used for post-processing are coded within a specialized branch of the software SPUX (Šukys and Bacci, 2021), and partly rely on the R scripting language (R Core Team, 2020).

3. Results and interpretation

According to the methodology recommended by Reichert et al. (2021), we first check convergence of the posterior sampling process (Section 3.1), then analyze potential problems of misuse of stochastic degrees of freedom to compensate for model structure deficits and interpret these deficits (Section 3.2), and finally analyze and discuss selected results to emphasize our findings (Section 3.3).

3.1. Convergence of sampling procedure for Bayesian inference

Convergence for deterministic models is achieved easily, even when we infer the exponent β of the storage-discharge relationship, which we do in a preliminary diagnostic run, see Fig. S1. In that case, a strong positive correlation is apparent between β and S_{uMax}, see Fig. S2, which is confirmed for model $Sto-D$, Fig. S3. For this model, a secondary posterior mode for small values of β and S_{uMax}, and large values of C_e, exacerbates convergence issues, Fig. S4. Hence, correlation problems and small secondary modes, both appearing when inferring STD parameters and β concurrently, let us lean towards fixing $\beta = 5$. This value roughly corresponds to the marginal posterior mode of model *Det*, Fig. S1.

Figs. S5–S9 show the evolution of the Markov chains for all the models when β is set to 5. Our core results are based on these models, although we have also performed additional control simulations with β fixed to 1 or 2, which we do not find to substantially alter the overall picture that we describe below. With $\beta = 5$, we notice that secondary modes are eliminated when a large initial portion of the data is discarded, see in particular Figs. S5b–S9b. In some cases, for instance for parameters D and C_e in model $Sto-K_s$, see Fig. S7, or parameters $\mu^{OU}_{logit(D)}$, k_s and $\sigma^{OU}_{logit(D)}$ in model $Sto-D$, Fig. S8, it is clear that at least 2500 parameters batches are required for burn-in not to surely falsify convergence. These correspond to 100 k parameters samples as we propagate 40 chains in parallel. Hence, if not differently specified, our results are based on the last 450 batches (18 k parameters samples).

The achieved slow convergence underscores the heavy computational load and wall-clock time required by this type of investigations. Overall, to conduct the numerical experiments, we use about 75 k node-hours on the Swiss National Supercomputing Centre's flagship machine Piz-Daint using on average 180 parallel processes per stochastic model, roughly partitioned 3:1 between the parallel propagation of the Markov chains and the parallel propagation of the hydrological and stochastic models. Memory limitations bound the maximum number of samples that we can analyze to $\sim$20 k. Tackling this limitation is dealt with in a newer version of our software (Šukys and Bacci, 2021).

Despite this large sampling effort, in the worst case scenarios ($Sto-K_f$ and $Sto-K_s$) we can not obtain more than a few hundred independent samples, see Table S2. Figs. S10–S14 quantify the overlap among Markov chains. This is done by comparing one another the $2.5-97.5$ percentiles spanned by the Markov chains in parameters space, see Figs. S10a,b-S14a,b. For the most critical cases, we also provide the evolution of the chains with the smallest overlaps for specific parameters, Figs. 10c,d–S14c,d. Taking these results together, we deem the overlap among chains acceptable, although we cannot fully dispel concerns regarding possible within-mode mixing barriers for a few parameters.

It is well-known that the PF method can degenerate due to high levels of stochasticity, high-dimensionality of the state space, and/or extreme observations. Filter degeneracy usually implies a poor estimate of the marginal likelihood, which in turn can hamper the estimation of the posterior density. To consider this aspect, we monitor the number of independent particles present in the filter at each resampling step. As shown in Figs. S15–S16, there are some cases when the filter collapses onto just one or two particles. These episodes are mainly limited to specific observational times. However, at those critical time-points up to

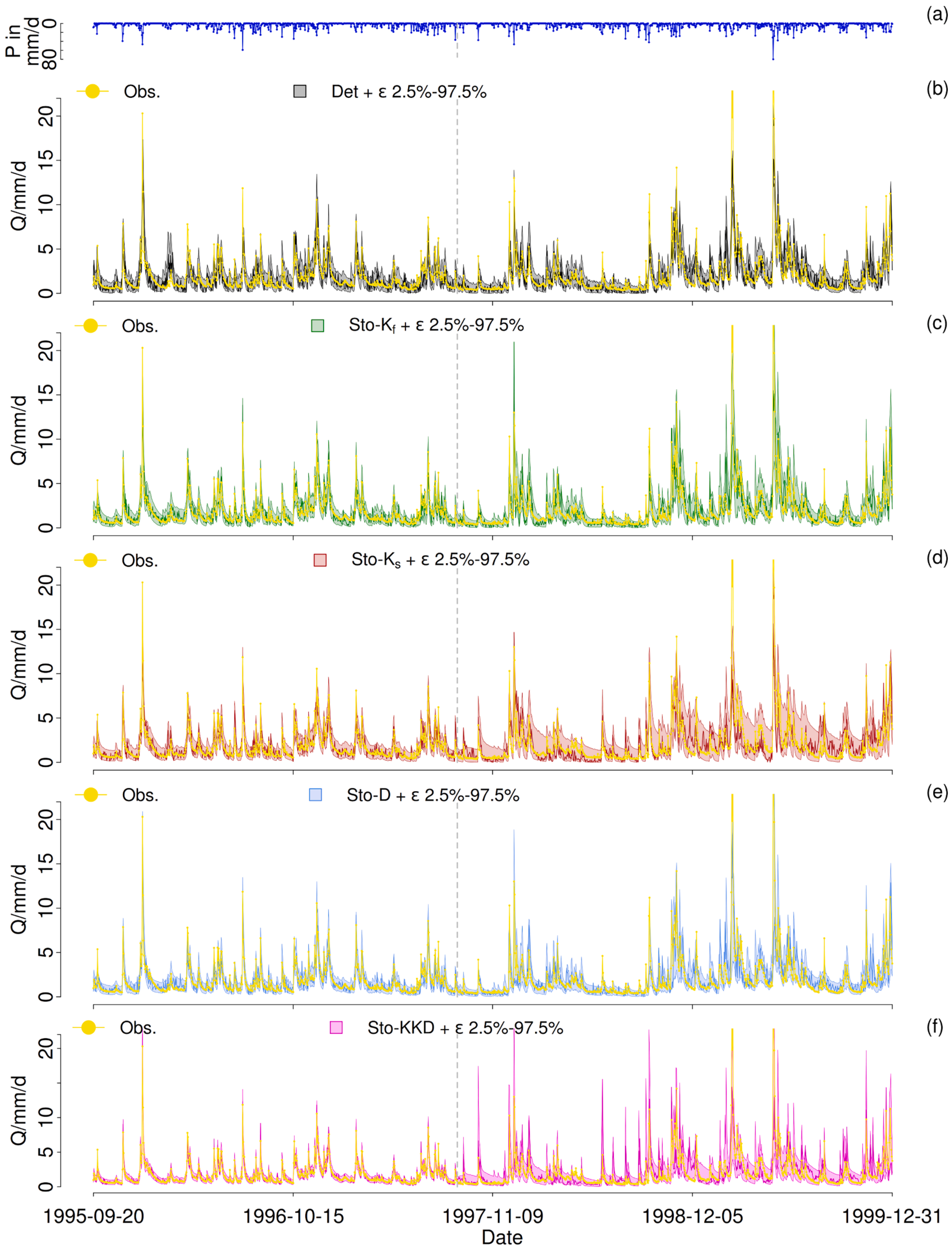


Fig. 3. Outflow Q for end of calibration and for cross-validation. Vertical dotted lines separate the last ~1 year of inference from cross-validation. Lines depict single model realization 2.5 –97.5% uncertainty bands. **(a)** Precipitation. **(b)** Outflow plus output error for model *Det*. **(c)** Same as (b) for $Sto-K_f$. **(d)** Same as (b) for $Sto-K_s$. **(e)** Same as (b) for $Sto-D$. **(f)** Same as (b) for $Sto-KKD$.

about 50% of the observational likelihood evaluations for models $Sto-K_f$ or model $Sto-KKD$ can result from collapsed filters. Despite these unsound cases account for only a very small fraction of the total number of observational likelihood evaluations, we also consider a diagnostic run for model $Sto-K_f$ where we double the number of particles (from 24 to 48). This inference produces results that are very similar to the ones relevant to 24 particles, see marginal posteriors in Fig. S17. Although this is not enough to completely rule out possible detrimental effects owed to filter collapse, it is a clear indication that the confined collapses of the filter that we experience are unlikely to seriously hinder our results. However, it is also clear that filter collapses can in general be critical for this type of applications. Attempts to use, during a preliminary inference phase, a number of particles that varies according to a measure of quality on the estimate of the marginal likelihood is a recent addition to our software framework. Incidentally but importantly, the results in Fig. S17 provide us with a strong and positive assessment of convergence.

3.2. Deficit analysis

3.2.1. Cross-validation

Visual Assessments. Visual inspection indicates a clear difference between the selected models regarding predictions of Q, see Fig. 3. Models Det, $Sto-K_f$, and $Sto-D$ appear to generate a similar extent of total predictive uncertainty in calibration and cross-validation, in the sense that there is a seamless transition between these two periods. On the contrary, model $Sto-K_s$ seems to produce larger uncertainty bands as soon as the filter is switched off, that is, as soon as cross-validation starts. A similar but less dramatic behavior is visible for model $Sto-KKD$ as well, which also suggests overconfidence during calibration when compared to all other cases. In terms of relative contribution of parametric and residual uncertainty to model predictions, as apparent when comparing the modeled output data with and without the contribution of the error term, see Figs. 3 and S18 respectively, the models show distinct behaviors based on their type (deterministic or stochastic). The predictive uncertainty of model Det is almost exclusively ascribable to the lumped error model. In contrast, the quota of uncertainty due to the observational noise is much diminished for all the stochastic models. We expect that these visual differences be mirrored in and quantified by those metrics that summarize the performance of a model. This comes next.

Quantitative assessments

The perceptions illustrated above are quantified by first considering the relative spread, Eq. (13). As mentioned in Section 3.2, we explore different possible aggregations of the data, and decide to plot in Fig. 4 the values of the spread when aggregating by particles, Eq. (13b), and the single scalar from aggregating all data per time, Eq. (13a), as aggregation by parameters gives very similar results to aggregating parameters and particles concurrently (tight distributions close to the scalar value).

Fig. 4 shows that for model Det, the spread of model predictions attributable to parametric uncertainty alone is the smallest both during calibration and cross-validation (left column in Fig. 4). For this model, predictive uncertainty is accounted for almost entirely by the residual error term (right column in Fig. 4).

For the stochastic models, three observations are plain. First, the spreads of model predictions attributable to parametric uncertainty are much larger than for model Det, and accounts for a substantial part of the total predictive uncertainty (roughly half, when comparing the left and right columns of Fig. 4). Second, spreads are smaller in inference than in prediction in contrast to model Det. Both results are evident indications of an improved partitioning of uncertainty by the stochastic models, which better reflects the different knowledge status about the system during calibration (more certain as conditioned on the observations) and cross-validation (more uncertain as the observations during the cross-validation period are not used to derive prediction uncertainty). Third, models $Sto-K_s$ and $Sto-KDD$ have an even much wider uncertainty during cross-validation than in calibration compared to the models $Sto-K_f$ and $Sto-D$, indicating a potential misuse of the stochasticity of model parameters during inference. This does not seem to be the case for $Sto-K_f$ and $Sto-D$, as their σ_{rel} is slightly lower than the one of model Det in prediction, Fig. 4d. Additionally, overconfidence for model $Sto-KKD$ in calibration is manifest in the relevant distribution of σ_{rel} in Fig. 4b.

Fig. 5 shows the evaluations of the cumulative distribution function (CDF) of the model predictions at the observational data points, see eq. (14). If our predictions would be perfect, the obtained CDF values of Q_{obs} should be uniformly distributed. In inference, the CDF metric spans all scenarios, from peaked tails for models Det and $Sto-D$, Fig. 5a, to excessive data-fidelity for $Sto-K_s$ and $Sto-KKD$, Figs. 5a,b (see also Figs. 4a,b to indeed assess that the respective σ_{rel} is small). In these cases, the more ideal scenarios seem to pertain to model $Sto-K_f$ without

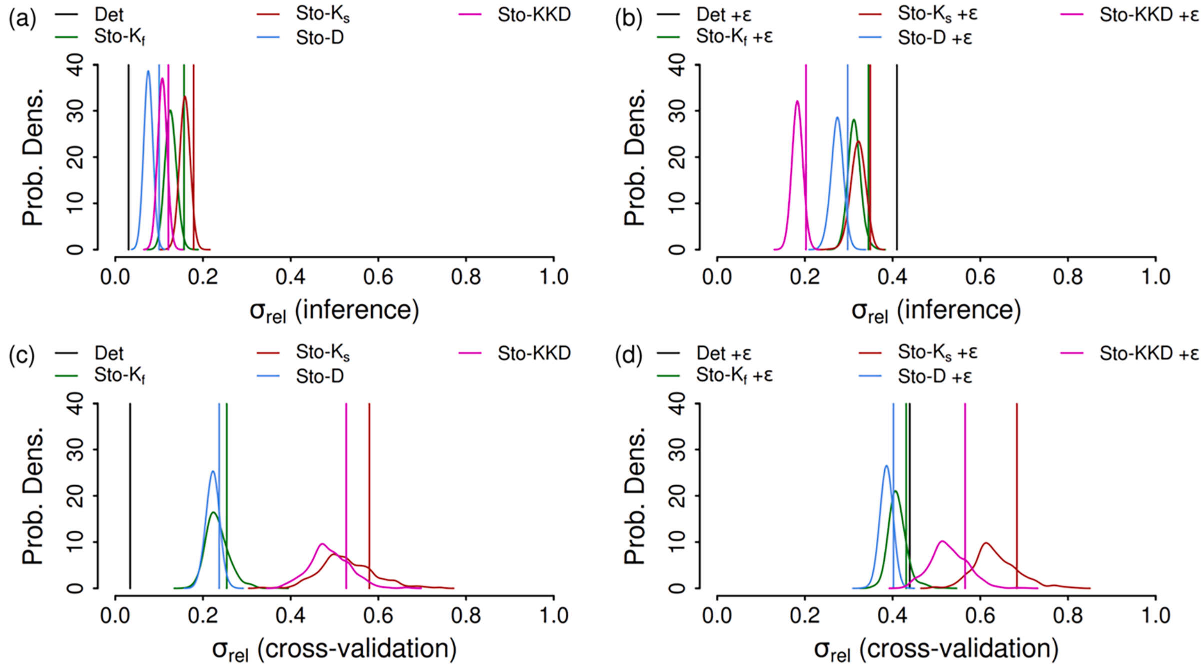


Fig. 4. Spread for Q with and without error terms from inference or cross-validation. Densities are obtained by a tight kernel density estimation and refer to data aggregated across particles, Eq. (13b). Vertical lines indicate the scalar values when spreads are calculated by aggregating both particles and parameters (this is the only value available for deterministic simulations), Eq. (13a). **(a)** Spread for Q from inference simulations and without errors. **(b)** Same as (a) but with errors on top of model output. **(c)** Same as (a) but from cross-validation simulations. **(d)** Same as (a) but from cross-validation simulations and with errors.

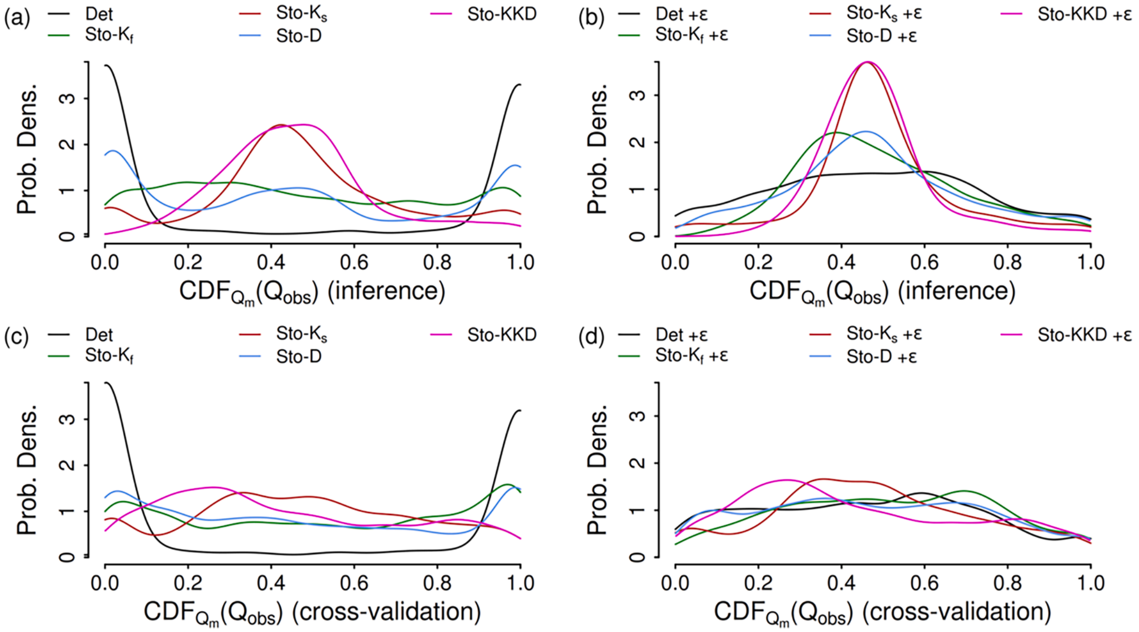


Fig. 5. CDF value for Q_{obs} with and without output errors from inference or cross-validation. Densities are obtained by kernel density estimation and refer to data aggregated across particles and parameters, Eq. (14). **(a)** CDF values of Q_{obs} when the CDF is computed from model's inference simulations without output errors. **(b)** Same as (a) but with output errors on top of model output. **(c)** Same as (a) but for cross-validation simulations. **(d)** Same as (a) but for cross-validation simulations and with output errors added to the model output.

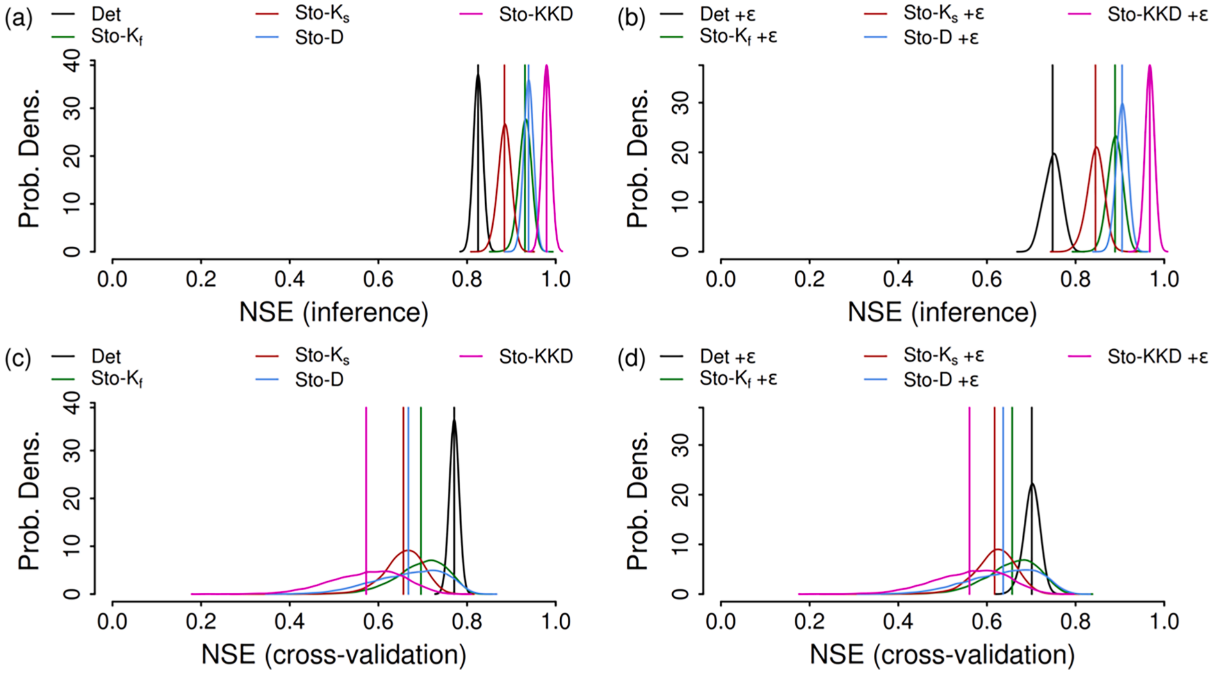


Fig. 6. NSE with and without output errors for inference and cross-validation. Vertical lines refers to the single scalar obtained by taking the expected value across the NSE values gained trajectory by trajectory. Kernel density estimates refer to the distribution of those values. **(a)** NSE computed from Q without output errors and from inference simulations. **(b)** Same as (a) but with output errors on top of model output. **(c)** Same as (a) but from cross-validation simulations.**(d)** Same as (a) but with output errors and from cross-validation.

observational noise and to model *Det* with output error. However, when we consider cross-validation with errors, Fig. 5d, all models perform comparably well. Overall, the CDF data underscore under a different corner the crucial role of the error model for model *Det* as opposed to the very small role of parametric uncertainty, while stochastic models seem confirmed to better partition the uncertainty into the different sources, providing adequate scenarios under most circumstances, especially in prediction, Figs. 5c,d.

Besides considering spread and CDF values, we are interested in assessing model performance through customary hydrological metrics. Fig. 6 reports the values of the Nash–Sutcliffe efficiency (NSE). As described in Section 2.8, these metrics are computed trajectory by trajectory, and reported using the full distributions and the associated expected values. The NSE decreases from inference to cross-validation (top

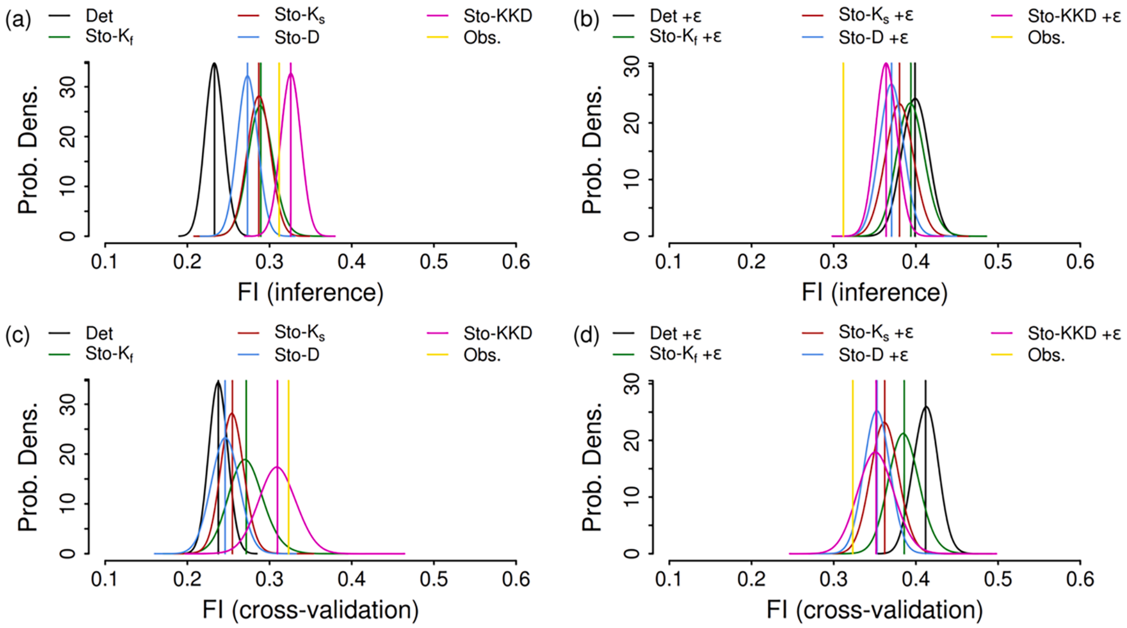


Fig. 7. FI with and without output errors for inference and cross-validation. Vertical lines refer to the single scalar obtained by taking the expected value across the FI values obtained trajectory by trajectory. Kernel density estimates refer to the distribution of those values. **(a)** FI computed from Q_m without output errors and from inference simulations. **(b)** Same as (a) but with output errors on top of model output. **(c)** Same as (a) but from cross-validation simulations. **(d)** Same as (a) but with output errors and from cross-validation.

vs bottom row of Fig. 6). This drop is larger for the stochastic models than for the deterministic one because of the better fit during calibration and the higher uncertainty in prediction that also deteriorates the mean. For the same reason, NSE also generally decreases when adding the lumped or observational uncertainty (left vs right column of Fig. 6). In this case, it is the stochastic models that minimize the degradation in performance. This result is a consequence of the observational noise playing a less important role in the stochastic models than the lumped error in the deterministic model. Among the stochastic models, $Sto-KKD$ and $Sto-K_s$ are the ones performing worst during cross-validation. In particular, while $Sto-KKD$ excels in calibration, it is also the model that shows the smallest NSE values in cross-validation. These results bring into question the assumptions of these models, such as an inappropriate selection of time dependent parameters. Note that it is not easy to distinguish a decrease in NSE during cross-validation due to model overparameterization or misuse of the stochastic degrees of freedom (which we try to avoid) from the decrease due to increased prediction uncertainty (which should be a natural feature of uncertainty quantification). The distributions of the NSE shown in Fig. 6 provide here an important insight. Fig. 6d shows that the right tails of the NSE distributions of models $Sto-K_f$ and $Sto-D$ extend to the largest values, even larger than those of the deterministic model, despite the mean NSE is smaller, lending this way additional confidence on the adequacy of models $Sto-K_f$ and $Sto-D$. On the other hand, this is not true for models $Sto-K_s$ and $Sto-KKD$, which is another indication of their inappropriateness.

Fig. 7 shows the assessment of model performance in terms of the flashiness index (FI). It is striking to see the influence of the lumped error on the FI for the *Det* model, with values that roughly double, moving from substantially below to substantially above the observed value. The stochastic models show an improved ability to match this signature. For those models, the FI does not dramatically increase due to the contribution of the observational error and, in any case, the FI of the stochastic models matches the observed value more closely than model *Det*. Additionally, FI values remain comparable across inference and cross-validation, similarly to what happens with model *Det*.

Taken together, these results indicate an improved partitioning of the uncertainty for the stochastic models $Sto-K_f$ and $Sto-D$ with respect to the deterministic case. Both of these models show the expected increase in uncertainty for prediction compared to calibration, an aspect that is missing for the deterministic model with a simple lumped error term. The nearly perfect fit of models $Sto-K_s$ and $Sto-KKD$ during calibration and worse performance than any other model during validation indicate an issue with overparameterization, or misuse of the stochastic degrees of freedom to compensate for model deficits during calibration, which cannot be replicated during cross-validation. We inspect the cause for these problems by analyzing the time series of the STD parameters and potential correlation with states and external influence factors in the next section.

3.2.2. Analysis of posterior parameter time series

When using stochastic models it is important to ensure that the stochastic processes are not compensating for substantial structural model errors, as in that case statistical assumptions would be violated. As concluded in the previous section, we expect such problems in particular for models $Sto-K_s$ and $Sto-KKD$. To scan for those, we first look at 2D projections that relate model inputs and states to the dynamics of the STD parameters, see Figs. S21–S26. By visual inspection, we do not detect any strong trend between an STD parameter and a model's input or state variable. This is an indication that the dynamics of the STD parameters are not systematically compensating in an obvious way for a trivial and strong (but missing in the model formulation) deterministic relationship. However, given the results of the calibration, we also consider specific plots for better discerning to what extent the stochastic dynamics of k_s in model $Sto-K_s$ drives the discharge, especially the larger values of the Q_s component of Q, see also Fig. S27. While we do perceive a weak trend from a zero-correlation scenario, indicating a possible role of k_s in driving larger Q_s, this trend is also present in the projection against S_s, as it is also confirmed by looking at time series.

Fig. 8 compares the times series for Q, S_s, S_f, and the respective STD parameter for models $Sto-K_f$ and $Sto-K_s$, while keeping the deterministic model as a baseline for a specific and interesting phase of the calibration (see Fig. S28 for the whole time-series). Fig. 8 let us identify two issues. First, the parameter k_f increases in at least two of the

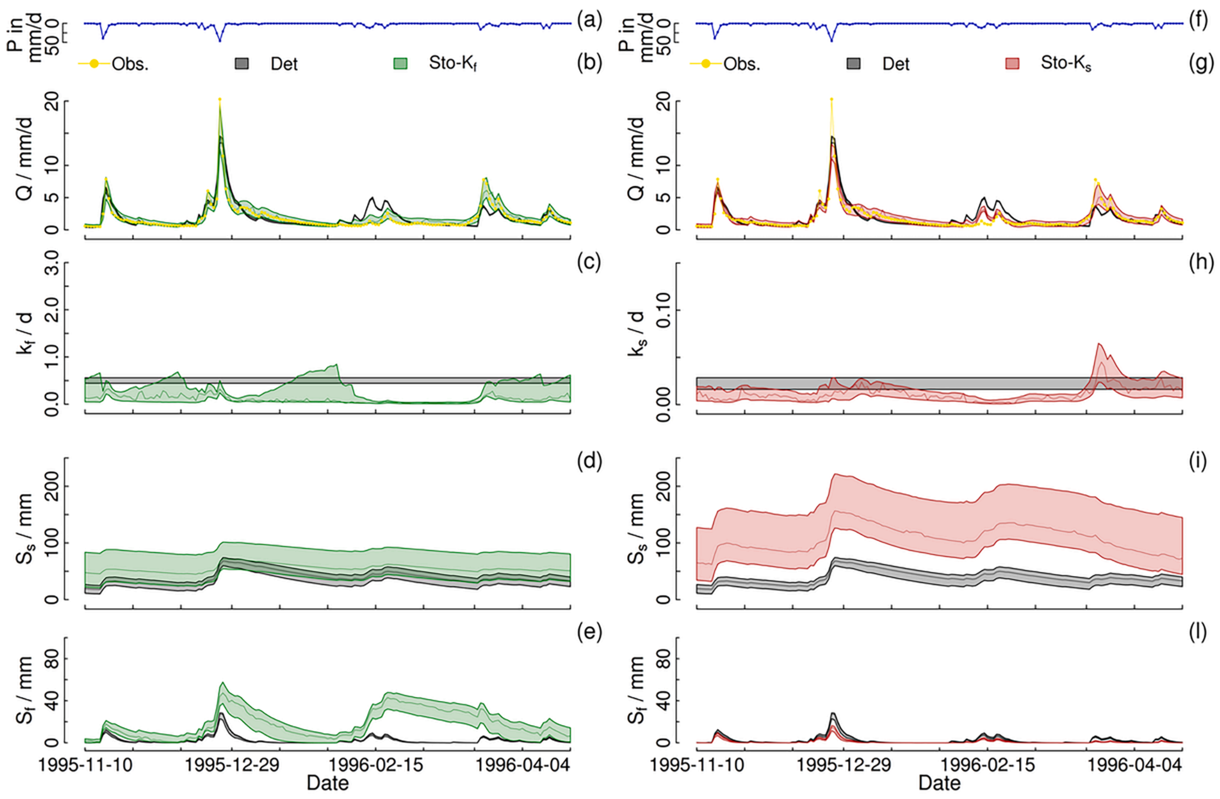


Fig. 8. Time series of model states and STD parameters. Inference, Nov 1995 - Apr 1996, no errors displayed, 2.5–97.5 percentiles. **(a)** Observed rainfall. **(b)** Trend of streamflow Q for models *Det* and $Sto-K_f$. **(c)** Trend of STD parameter k_f for *Det* and $Sto-K_f$. **(d)** Trend of reservoir level S_s for *Det* and $Sto-K_f$. **(e)** Trend of reservoir level S_f for *Det* and $Sto-K_f$. **(f)** Same as (a). **(g)** Same as (b) for models *Det* and $Sto-K_s$. **(h)** Same as (c) for *Det* and $Sto-K_s$. **(i)** Same as (d) for *Det* and $Sto-K_s$. **(l)** Same as (e) for *Det* and $Sto-K_s$.

recession phases. This indicates that a nonlinear release relationship of the fast reservoir may reduce deficits of the deterministic model. The time series of k_s let us identify the problem of model $Sto-K_s$. It becomes evident that a fast, systematic variation of this parameter considerably contributes to the generation of the fourth discharge peak. It is a general danger of adding stochasticity to a release parameter of a slow reservoir that during calibration the variation of this parameter can be misused to generate any desired outflow, as this reservoir will hardly ever be empty. The reason why this seems to happen for just the fourth discharge peak in the selected calibration period can be seen from the deterministic simulation in Fig. 8g. The deterministic model underestimates this peak. A possible cause is thus that for this event the input was underestimated probably because the storm only partly hit the rain gauge. However, the fact that the misuse of k_s occurs interspersed by increases in the water level S_s, which also drives Q_s higher (see Figs. S27–S28), makes this problem hardly identifiable from the simple scatter plots discussed above. Indeed, Q_s can be large also when k_s is comparably small. Hence, a 3D analysis of Q_s, k_s and S_s would be optimal in this case, albeit difficult to visualize and anticipate. This demonstrates the importance of the cross-validation analysis done in Section 3.2.1 to identify this kind of problems. The poor behavior of the model $Sto-KKD$ just results as a consequence that it also contains the STD parameter k_s. As we expect stochastic variables of a release coefficient of a slow reservoir to be also slow, the question remains whether it would have been possible to avoid this behavior by a stronger prior for the correlation time of the STD parameter k_s. This will be investigated in the next subsection.

3.2.3. *Control simulation for model $Sto-K_s$*

In an attempt to establish if the apparent misuse of parameter k_s is actually just due to a lack of convergence owed to unfavorable starting conditions and/or naive prior belief for parameter $\tau^{OU}_{\ln(k_s)}$, we perform an extra simulation where we increase the mean of the prior of the correlation time from 12 to 60 days. Results appear well-converged as for $\tau^{OU}_{\ln(k_s)} = 12$, see Fig. S29. The log-posterior values depicted in Fig. S30 are comparable, suggesting a lack of clear preference for one of the two cases, which confirms the notion that the parameters of the OU process are notoriously difficult to identify. This is also supported by Fig. S31, which makes the point that different marginal posteriors for the parameters of the OU process do not necessarily imply fundamentally different marginal posteriors for the other parameters. However, it also provides us with a strong positive assessment on convergence of calibration for all the other parameters.

Unfortunately, due to a correlation between $\tau^{OU}_{\ln(k_s)}$ and $\sigma^{OU}_{\ln(k_s)}$, which was already present in the original simulations, see Fig. S32, by favoring larger values of $\tau^{OU}_{\ln(k_s)}$ we also favor larger fluctuations in the dynamics of k_s. These impede an improvement of the metrics with respect to what already shown in Section 3.2.1, see Figs. S33–S36, highlighting this way a non-trivial balance between overparameterization and improved description of our state of knowledge when using k_s as STD parameter, which might be difficult to control and optimize. Indeed, it appears that the misuse is also not corrected by the obtained larger values of the autocorrelation time, as from Fig. S37 we evince the striking similarity of the results irrespective of the different prior means that we use.

3.2.4. *Summary of deficit analysis*

In summary, from our deficit analysis, we conclude that we would have to exclude the models with the STD parameter k_s from posterior analysis and prediction. This decision could only be revised if we would either be able to provide better input data or include a dynamic input uncertainty model, such as the one used by Del Giudice et al. (2016).

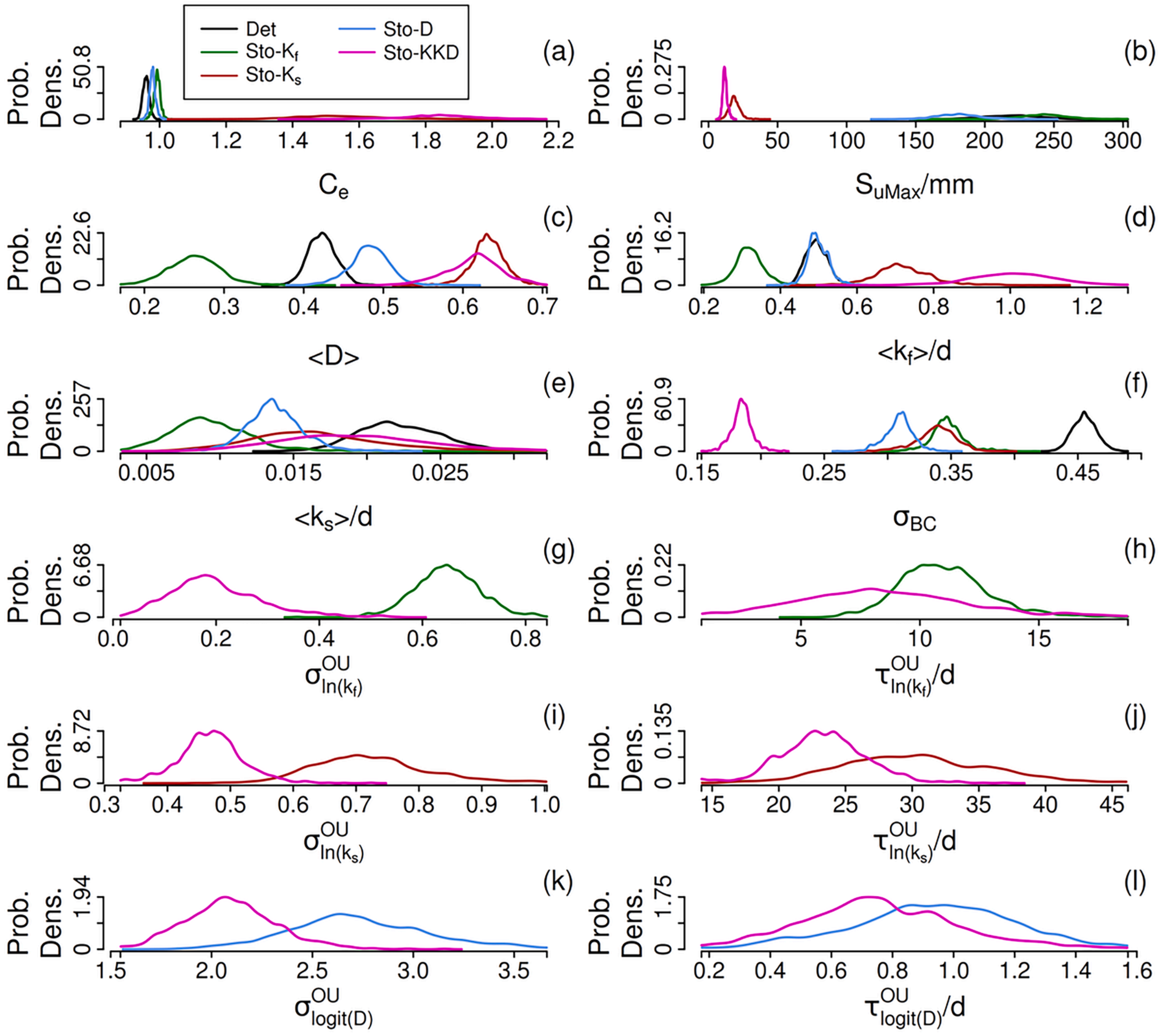


Fig. 9. Marginal posterior distributions for the inferred parameters. Distributions are clipped at 2.5%-97.5% percentiles. Stochastic parameters are displayed as transformed-back mean of the corresponding stochastic process, see Section 2.3. The legend in panel (a) applies to all panels.

3.3. Posterior analysis

As presented in Section 3.2, models $Sto-KKD$ and $Sto-K_s$ are performing poorly compared to the other models and should be discarded. However, we keep them in the following analysis to didactically point out some additional features that the misuse of a stochastic parameter can cause.

3.3.1. Parameters marginal posteriors

Fig. 9 shows the marginal posterior distributions of the parameters of the various models. The most striking differences are apparent for parameters C_e and S_{uMax}. Models $Det, Sto-K_f$, and $Sto-D$ behave similarly by showing peaked distributions for C_e and corresponding large uncertain values for S_{uMax}. Models $Sto-K_s$ and $Sto-KKD$ show the opposite behavior, with wide distributions for C_e shifted to very large values, and peaked small values for S_{uMax}, see Figs. 9a,b, which are likely just another footprint of the misuse of STD parameters. Incidentally, the obtained values of the evaporation parameter C_e for the models with stochastic k_s clearly point to problems as they are much larger than expert knowledge would suggest. Additionally, models $Sto-K_s$ and $Sto-KKD$ show a correlation between C_e and S_{uMax}, which is not present in any other model, see Fig. S19. This difference in behavior can be interpreted considering that models $Sto-K_s$ and $Sto-KKD$ have much smaller values of S_{uMax} than models $Sto-K_f$ and $Sto-D$. The values of S_{uMax} for models $Sto-K_s$ and $Sto-KKD$ may even approach zero, meaning that the reservoir can run empty and cannot evaporate. The closure of the water balance in these models is then achieved by increasing the evaporation to unrealistically high values (i.e. C_e values much larger than one) when water is available in the reservoir.

For what concerns the splitting parameter D, models can be visually (and approximately) separated into three groups, see Fig. 9c. In model $Sto-K_f$ parameter D tends to be smaller than 0.5, which favors routing the precipitation to the fast reservoir. In contrast, inference for $Sto-K_s$ and $Sto-KKD$ results in $D > 0.5$. Models $Sto-D$ and Det do not seem to favor the routing of the rainfall to a specific reservoir. This result indicates that a stochastic dynamics of a release coefficient tends to foster higher fluxes of water to the corresponding reservoir, minimizing the flux to the others. In our study, this has particularly detrimental consequences when the reservoir in question is the slow one, see Section 3.2.1.

Another aspect of marginal posteriors that is worth to point out is relevant to parameters $\tau^{OU}_{\ln(k_f)}$ and $\tau^{OU}_{\ln(k_s)}$. The difference in these time scales is limited to just about a factor of 3 (32 vs 11 days at mode values). Despite we expect a larger value for the correlation time of the OU process for k_s than the one for k_f, the obtained difference is not enough, as we see in Section 3.2.2, to avoid response of the slow reservoir to quite fast outflow events for models *Sto* $-K_s$ and *Sto* $-KKD$. Unfortunately, we also establish in Section 3.2.3 that it is difficult to correct this behavior straightforwardly, as it is not enough to just impose a lager prior value for $\tau^{OU}_{\ln(k_s)}$, as parameters posterior distributions differ only for $\tau^{OU}_{\ln(k_s)}$ and $\sigma^{OU}_{\ln(k_s)}$, while caveats persist. Although $\tau^{OU}_{\ln(k_f)}$ is also not small enough to completely avoid that the fast reservoir contributes to the recession leg, it is clear that this could be corrected by a non-linear relationship, and that this problem is much less severe, as it does not undermine the model's performances appreciably, see Section 3.2.1.

3.3.2. States marginal posteriors

In Fig. S20 we group together the values of some of the variables that compose the state of the models as resulting from calibration simulations. Those include the outflow Q_m, the fluxes within the model Q_f and Q_s, the level of the water in the reservoirs S_u, S_f and S_s, and the value of the STD parameters, if present. The differences in the marginal posteriors of the parameters in Fig. 9 are mirrored in the marginal posterior of the states.

Models *Sto* $-K_s$ and *Sto* $-KKD$ differ from the other models especially in the distribution of S_u, which is depleted of the bulk values, resulting in an often almost empty unsaturated reservoir. The distribution of S_f also appears thinner. Hence, evaporation and the level of water in the slow reservoir have to compensate. This is just one extra confirmation that when k_s is made stochastic, inference can promote a model where the slow response leads not only the baseflow, but also part of the fast dynamics, and this reciprocates with large values of C_e and small values of S_{uMax}. The distribution of Q_s also contains the footprint of the reaction of the slow reservoir to fast dynamics. Necessarily, this reduces the number of peak flows from the fast reservoir, which is a consequence visible in the distribution of Q_f for model *Sto* $-K_s$, see Fig. S20c. A diminished role of the fast reservoir is also indirectly noticeable in the distribution of D in model *Sto* $-KKD$ when compared to the same distribution in model *Sto* $-D$, see Fig. S20d,e. Other repercussions of the different underlying model behavior when a release coefficient is made stochastic are

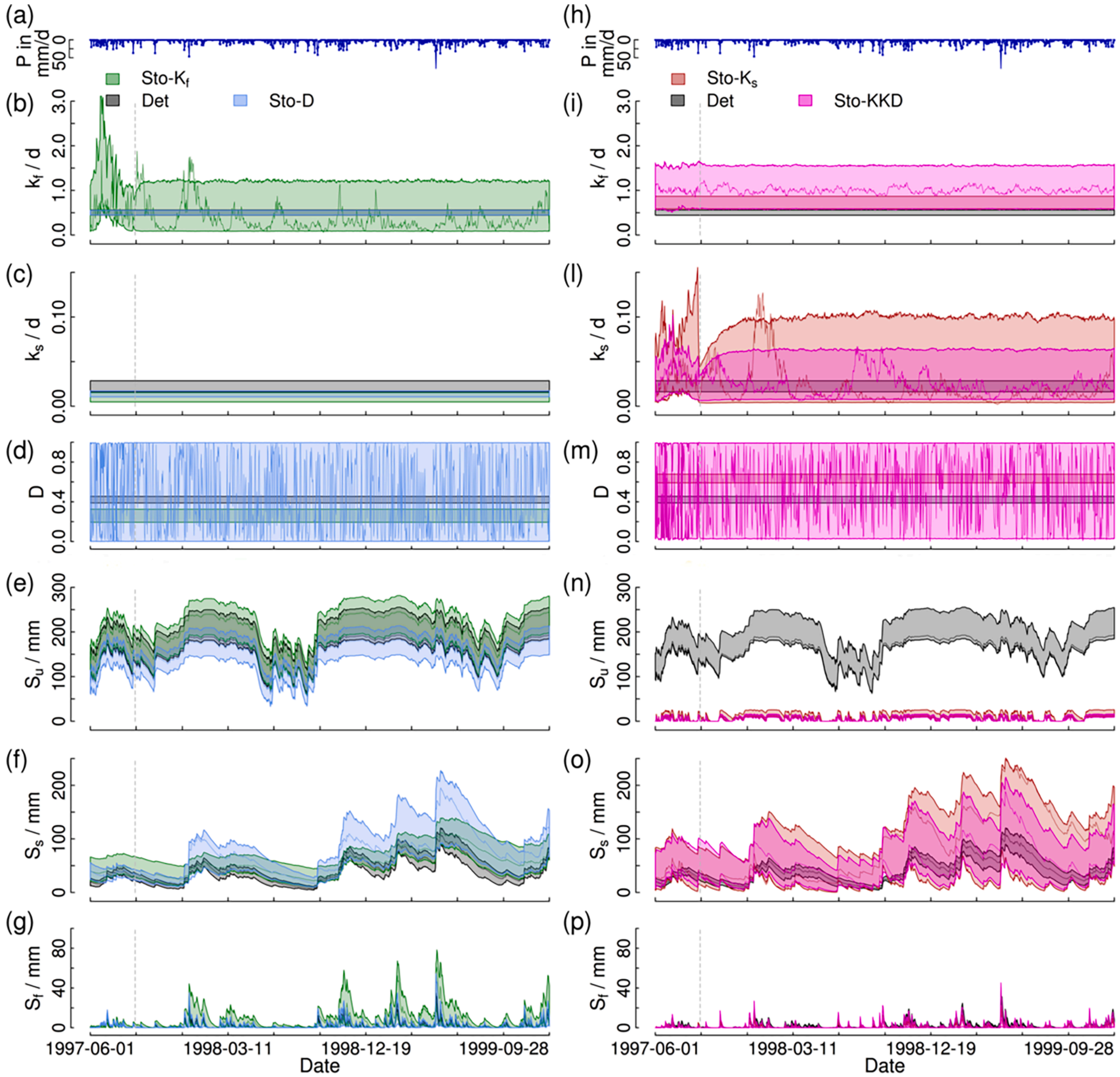


Fig. 10. Trends of STD parameters and water levels at end of inference and during cross-validation. **(a)** Observed rainfall. **(b)** Parameter k_f for models *Det*, *Sto* $-K_f$, and D **(c)** Same as (b) for parameter k_s. **(d)** Same as (b) for parameter D. **(e)** Water level in the unsaturated reservoir for models *Det*, *Sto* $-K_f$, and D. **(f)** Same as (e) for the slow reservoir. **(g)** Same as (e) for the fast reservoir. **(h)** Same as (a) **(i)** Same as (b) for models *Det*, *Sto* $-K_s$, and *Sto* $-KKD$. **(l)** Same as (i) for parameter k_s. **(m)** Same as (i) for parameter k_s. **(n)** Same as (e) for models *Det*, *Sto* $-K_s$, and *Sto* $-KKD$. **(o)** Same as (n) for the slow reservoir. **(p)** Same as (n) for the fast reservoir.

apparent when we compare the distribution of k_f in model $Sto - K_f$ with the one of the other models. An increased role of the fast reservoir is strongly suggested by the distribution in Fig. S20b. Importantly, these considerations are much easier in hindsight after cross-validation, see Sections 3.2.1 and 3.2.2.

3.3.3. *Posterior predictions*

Fig. 10 illustrates the dynamics of the STD parameters and of the level of water in the reservoirs in the deterministic and stochastic models for the final part of the calibration period and for the whole validation period. These results complement and extend the analysis of the outflow Q already shown in Fig. 3 by focusing primarily on the cross-validation period and on the internal variables.

Overall, models $Det, Sto - K_f$ and $Sto - D$ produce similar dynamics throughout, both in terms of trends and values of the state variables of the process model. However, we should notice that there are specific differences within the $Det, Sto - K_f, Sto - D$ group in the trends for the level of water, especially for S_s and S_f, while S_u shows the smaller variability. It is hence interesting to point out that S_u looks completely different when k_s is stochastic, compare Figs. 10e,n, that S_s from model $Sto - D$ appears gaining a long-term dynamics more similar to $Sto - K_s$ and $Sto - KDD$, see Figs. 10f,o, and that the dynamics of S_f for $Sto - K_f$ gains momentum at specific time-points well above what happens for the Det and $Sto - D$ models.

4. Discussion

After having shown and interpreted the results of our case study, here we discuss potentially generalizable new results. It is appropriate to remark that our contribution centers on investigating and discussing the improvements that we can obtain in characterizing our uncertain knowledge about a hydrological system when we resort to using stochastic time-dependent parameters. Hence, we do not necessarily search with our approach improvements to hydrological peak perfomance quantified by goodness of fit metrics such as the NSE. In fact, the deterministic model that we employ already possesses satisfactory skills in reproducing the data according to this metric. Rather, it is its ability in partitioning the different sources of uncertainty and in describing the status of knowledge, which clearly differs from calibration to validation, that we want to improve upon. By doing so, we are also able to conceive data-driven suggestions for the improvement of the process representation in the *Det* hydrological model for future assessments, which could allow maximizing peak performance while still including a stochastic representation of internal/intrinsic variability for its aptly partitioning between the different sources. In addition to improvements in the description of variability, describing uncertainties in hydrological models intrinsically, rather than imposing their overall effect on observed output at the end of the cascade of processes transforming input into output, is also more appealing conceptually, as previously discussed and demonstrated with didactical models (Reichert et al., 2021). Compared to these previous examples, our more realistic conceptual model applications confirm the appropriateness of the proposed method in improving the partition of variability, and lead to additional insights on how different modeling choices may affect the performance of the stochastic models, occasionally leading to an unintended behavior, but also possibly exposing routes to model improvement. From our results, we expect an improved description of uncertainty of the model with STD parameters compared with the model with constant parameters and a lumped error term if we are able to choose stochastic parameters that do not lead to a systematic correction of deficits of the deterministic model during calibration. Recommendations on suitable diagnostic and data-mining approaches, as well as on suitable modeling choices are discussed next.

A key difference in behavior between intrinsically stochastic models and deterministic models with a lumped error term is represented by their predictive performance in calibration and cross-validation. Stochastic models during calibration can have low predictive uncertainty and achieve an excellent fit to the observations, due to the calibrated time courses of the time dependent parameters. During validation, the uninformed (due to lack of data-awareness) stochastic variability tend to lead to larger predictive uncertainties than in calibration and, in case of misuse of the stochastic degrees of freedom in calibration, even to a critical degradation of performances. Deterministic models with a simple lumped error term, in contrast, due to their fixed parameters, do not tend to show a significant difference in predictive uncertainty between calibration and cross-validation.

The fact that stochastic models lead to a higher uncertainty in prediction than in calibration is arguably an appealing feature of these models, as it reflects that during calibration the data are known, whereas during prediction the knowledge description by the posterior probability distribution is based on the data during the calibration period. The absence of this behavior in deterministic models with a simple lumped error term can be interpreted as a symptom of unrealistic behavior, and in particular of their inability to distinguish between the conditions of whether the data are known. However, when dealing with stochastic models, it is important to assess when such an increase in uncertainty in model predictions from calibration to cross-validation is desirable, or at least unavoidable, from when it is excessive and symptomatic of overfitting during calibration. In order to distinguish between these two cases, a careful analysis of model performance during cross validation is essential. Therefore, cross-validation plays a crucial role in the evaluation of stochastic models, much more than for deterministic models, where the performance during calibration and cross-validation can be very similar. Our analyses indicate that in order to operate this distinction, it may be useful to inspect the posterior distributions of the NSE during cross-validation. The mean NSE will drop in the prediction phase compared to the calibration phase due to the increased uncertainty (that also allows for smaller values). This is similar to overfitting with the mechanistic model. However, a comparison of the right tail of the posterior NSE distribution could indicate which stochastic models produce better NSE values than the deterministic approach, and which do not. In our case study, this diagnostic leads to a clear separation of the "problematic" from the "realistic" stochastic parameters. To discover the possible cause(s) of misuse of STD parameters, we find it important to analyze the time series of the parameters themselves, and of the other variables that describe the internal state of the model. This can also allow realizing opportunities for model improvements in a data-driven fashion, especially for the "realistic" settings, as in our case it would be natural to propose an improved hydrological model where the water release from the fast reservoir is controlled by the water level through a non-linear relationship. However, more experience is certainly needed to confirm the value of these analyses, and to identify more indicators to support a good selection of stochastic parameters.

The differences between different stochastic models regarding producing a desirable behavior in prediction, raise the question of whether more general recommendations can be given on which model parameters are good candidates to reflect model intrinsic uncertainty. In our examples, bad performing models are associated with the choice of making release coefficients of slowly reacting reservoirs stochastic. This result suggests that this choice should be avoided. Slowly reacting reservoirs typically model groundwater and are of primary importance for modeling base flow. Consequently, such reservoirs are hardly ever empty. For this reason, with time-variation of its release coefficient, (nearly) any dynamic discharge pattern can be produced. This is dangerous because if the hydrologic model is unable to produce some observed pattern either due to input or model structural errors during model calibration, this release coefficient can be misused to produce this pattern. We have shown that, because of the amount of hydrological data, even a strong prior in favor of a long correlation time of such a release coefficient cannot avoid this problem. This is certainly an issue that would occur in other applications as well. It can be resolved by not making this coefficient stochastic, even if some stochasticity due to varying releases from different water bodies with different release behavior would make a slow and hard to predict temporal variation realistic.

In our case study, the choice of making the release coefficient of the slow reservoir stochastic has a detrimental effect also on other parameters. It leads, in fact, to higher water demand for the slow reservoir, which is achieved by modifying the water division coefficient and other model parameters. This results in unrealistic values of multiple constant model parameters of other model components. This is a phenomenon that would probably occur also in other applications. The conclusion is that unrealistic (in low-probability domains of the prior) posterior parameter values can be an additional indication of problems with stochastic parameters and need careful analysis.

In summary, our results indicate that some general instruments can be used to distinguish the expected higher uncertainty during prediction periods (as in cross-validation) from excess uncertainty: (i) the analysis of the distribution of predicted Nash Sutcliffe Efficiencies (NSE) during cross-validation, and (ii) the shift of constant model parameters beyond their prior high-probability range. As a general recommendation regarding the choice of time dependent parameters, we highlight the particular danger of making release coefficients of slowly reacting reservoirs stochastic. Additionally, we also highlight the opportunity inherent to the analysis of the time course of the STD parameters and related internal variables to act as a possible data-driven source for model improvement.

A final new aspect of our study is the application of a different numerical approach. Indeed, we apply a Particle Markov Chain Monte Carlo (PMCMC) approach as in Andrieu et al. (2010) to numerically sample from the posterior instead of conditional Ornstein–Uhlenbeck sampling as in earlier studies (Reichert and Mieleitner, 2009; Reichert et al., 2021). This is a general approach for sampling from stochastic state-space models. It combines a Particle Filtering (PF) process for sampling dynamic states (in our case stochastic parameters) and calculating an approximate marginal likelihood for use in an outer Markov Chain Monte Carlo (MCMC) procedure for sampling the constant parameters. The advantages of this approach are that the Monte Carlo algorithm can easily be parallelized, and that the PF does not require neither linearity in the process it samples from, differently from conditional Ornstein–Uhlenbeck sampling, nor normality in the distribution of the sampled space, differently from the Ensemble Kalman filter (Evensen, 2009). However, as any other approach to states estimation, it can suffer from all the caveats inherent to the sampling of high dimensional spaces (e.g., Verleysen and François, 2005), which are primarily manifest in filter collapses, see Section 3.1. Additionally, the particle resampling step poses challenges to the scalability of numerical codes. All these algorithms that we use are implemented in the recently developed framework SPUX (Šukys and Bacci, 2021).

5. Conclusions

Conceptual hydrological models are very successful in describing key features of observed discharge time series. On the other hand, they are highly simplified representations of streamflow generating processes, which leads to intrinsic model uncertainty that is propagated to the output. The traditional description of a hydrological system with a deterministic, conceptual model and a lumped output error model does not explicitly consider the main mechanisms of (intrinsic) uncertainty generation. Making mass fluxes between reservoirs stochastic by stochastic, time-dependent parameters is a means of describing such intrinsically generated uncertainty. The uncertainty in the states is then a consequence of the uncertainty in the mass fluxes. This is a conceptually more convincing concept than making mass-balance equations stochastic, because it is a closer description of the underlying mechanisms and maintains mass-balances exactly. On the other hand, statistical inference for stochastic models is methodologically and computationally much more challenging than for deterministic models with a lumped error term.

This study proposes a new implementation of stochastic, time-dependent parameters for Bayesian inference using a Particle Filter (PF) method coupled with a Markov Chain Monte Carlo approach. The method is tested on a real case study using a multi-reservoir hydrological model. In particular, we compare 4 stochastic hydrological model variants with different selections of time dependent parameters to a deterministic model variant. All variants include an “observational” error, which in the case of the deterministic model is intended to account for all sources of uncertainty. Our main conclusions are summarized as follows:

1. The combined Particle Filter, Markov chain Monte Carlo method provides a feasible alternative to previous implementations, with the advantages of a potentially more efficient inference procedure, and of a more general range of applications, as it does not mandate linearity and/or Gaussian assumptions. New inference frameworks, such as the SPUX framework tested in this paper, are meant to facilitate the application of the method and shorten the execution time by parallelization, albeit scalability is difficult to achieve.
2. The stochastic models have the potential to provide a more realistic description of uncertainty than the deterministic model. In particular, two out of the four stochastic models, namely $Sto-K_f$, which makes the water release rate parameter of the fast reservoir stochastic, and $Sto-D$, which makes the split parameter from the unsaturated to the fast and slow reservoirs stochastic, achieve a better description of our uncertain knowledge than the deterministic model with a lumped error term. This assessment is based on the following results: (i) Although the deterministic and the stochastic models have similar predictive uncertainty bands, the portion of this uncertainty attributable to parametric uncertainty is much larger for the stochastic than for the deterministic model, compare Figs. 3 to S18, and the left to the right column of Fig. 4. This appears more realistic, especially for the characterization of the states. (ii) Differently from the deterministic model, the uncertainty of model output is smaller during the calibration than during the prediction period for the stochastic models, which should be a natural result as it reflects our posterior knowledge of discharge given the observations during the calibration period, compare the top row of Fig. 4 with the bottom row. (iii) The stochastic models generate autocorrelated errors in output naturally either through the stochastic process (e.g. in the case of the water release rate parameter) or even combined with autocorrelation produced by downstream reservoirs (e.g. in the case of the split parameter). The effect of this behavior is manifested in the ability of the stochastic models to improve the match to hydrological signatures sensitive to autocorrelation, such as the flashiness index shown in Fig. 7. These appealing features of improved characterization of uncertainty and more faithful adherence of the flashiness index to the observed value, do not mar the other hydrological metrics for those stochastic models devoid of overparameterization effects. Indeed, we find that the NSE for models $Sto-K_f$ and $Sto-D$, while showing a larger variability during validation in compliance with the notion that the data are unknown, is not necessarily lower than the one of model *Det*, see Fig. 6 (in particular the right tail of the distributions). Similarly, the CDF metric does not seem impaired for those two stochastic models, see Fig. 5, and this despite a small reduction in the relative spread with respect to the *Det* model as in Fig. 4. All these observations taken together are an indication that a stochastic model can allow reducing the uncertainty in prediction for a similar value of consistency between the model output and the data in the validation set, while also improving the partition of the variability between the different sources and the modeling of the correlation effects.

3. Stochasticity can be misused to compensate for model or input deficits. This effect is shown by the other two stochastic models, namely *Sto* $-K_s$, which makes the water release rate parameter of the slow reservoir stochastic, and *Sto* $-KKD$, which makes all three considered parameters stochastic. These models have a significantly poorer performance during cross-validation than the other models, such as much larger uncertainty bands, and smaller NSE values. We attribute this effect to a misuse of time dependent parameters to compensate for model or input deficits. For example, model *Sto* $-K_s$ reproduces a discharge peak that could not be reproduced by the deterministic model through the time-variation of the water release rate parameter of the slow reservoir. This is hardly a realistic description of the underlying system. Such a behavior is difficult to avoid because time-dependence of the water-release rate parameter of a reservoir that usually has sufficient water content (as it is typical for a slow reservoir, such as a reservoir representing groundwater) can produce any dynamic output.
4. Our work indicates that model results can suggest additional modeling choices. In this context, looking at our results as a whole clearly suggests that making parameters D and k_f concurrently stochastic should appear as a natural choice for a future set of additional investigations. Similarly, diagnostic analyses for model *Sto* $-K_f$ turn into a data-driven discovery of a small model deficit that could be simply overcome by establishing a non-linear relationship for the dynamics of the fast reservoir. This could prove beneficial both for improving hydrological performance and for the description of parameteric and intrinsic variability.

The varying performance of stochastic models suggest that the choice of which model parameters are made time dependent is important. In this study, we find that it is challenging to add stochasticity to a slowly reacting reservoir, while we find more encouraging results for parameters linked to faster components of the model. However, insufficient experience is currently available to provide recommendations on which model parameters should be made stochastic.

Our work underscores many potential areas for future exploration:

- Consider the possible strategies for model improvement as identified through data-mining of the stochastic dynamics and of the influencing/influenced variables.
- Explicitly consider input error and higher time resolution in input data.
- Conduct more-in-depth sensitivity analyses on both prior and PF hyper-parameters (especially number of particles to avoid collapses completely).
- Perform more simulations on different case studies to improve our understanding about overparameterization dependence/effects on the individual parameters.
- Elucidate the dependence on/influence of the chosen stochastic process on the results.
- Investigate the sensitivity to the error model (BC vs. other approaches).

Regarding the suggested outlook, we would like to reiterate that investigating the effect of different input features, such as time resolution, and/or of an input error model are beyond the scope of this paper. This is due to the need to maintain computational time and budget, as well as data footprint, within available resources, and due to our decision to focus on analyzing the implications of choosing STD parameters among the ones of the hydrological process model. However, research relevant to the effects owed to the explicit modeling of the input error is currently in focus in our department.

Albeit we test the effect of doubling the number of particles, additional assessments on the consequences owed to filter collapses are hampered by the large scale investigations that we carry out. Indeed, here we focus on comparing multiple STD parameters, while we suggest that further diagnostic work shall focus on one specific case due to the computational load implied by changing the number of particles systematically.

Finally, further exploration of the effects of other modeling choices is of interest as well. At variance with conditional Ornstein–Uhlenbeck sampling, PMCMC would allow extension of our work to non-linear stochastic processes. This can be interesting for further research, as it seems plausible that the OU process can pose some limitations owed to correlations between $\sigma^{OU}_{f(\theta_s)}$ and $\tau^{OU}_{f(\theta_s)}$, as it is demonstrated by the diagnostic run where we change the prior of $\tau^{OU}_{ln(k_s)}$ with the aim to mitigate misuse. Similarly, different output error models can be object of further investigations too, as we find that by using a Box-Cox approach we are able to overcome identifiability issues relevant to the width of the error distribution (Reichert et al., 2021).

CRediT authorship contribution statement

Marco Bacci: Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Marco Dal Molin:** Software, Data curation, Writing - review & editing. **Fabrizio Fenicia:** Methodology, Software, Writing - review & editing. **Peter Reichert:** Conceptualization, Methodology, Writing - review & editing, Supervision, Project administration. **Jonas Šukys:** Software, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank Prof. Siddhartha Mishra (ETH Zurich) for access to the EULER cluster and Dr. Lorenz Ammann (WSL) for fruitful discussions. We also would like to thank the Swiss National Supercomputing Centre (CSCS) personnel for timely support and acknowledge that this work was supported by a grant from CSCS under project ID s950. We also would like to thank the Eawag Directorate for the financial support within the Discretionary Funding program.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jhydrol.2022.128057.

References

Ammann, L., Fenicia, F., Reichert, P., 2019. A likelihood framework for deterministic hydrological models and the importance of non-stationary autocorrelation. Hydrol. Earth Syst. Sci. 23 (4), 2147–2172.

Andrieu, C., Doucet, A., Holenstein, R., 2010. Particle Markov chain Monte Carlo methods. J. R. Stat. Soc.: Ser. B (Statistical Methodology) 72 (3), 269–342.

Baker, D.B., Richards, R.P., Loftus, T.T., Kramer, J.W., 2004. A new flashiness index: Characteristics and applications to midwestern rivers and streams 1. JAWRA J. Am. Water Resour. Assoc. 40 (2), 503–522.

Box, G.E., Cox, D.R., 1964. An analysis of transformations. J. Roy. Stat. Soc.: Ser. B (Methodol.) 26 (2), 211–243.

Buser, C.M., 2003. Differentialgleichungen mit zufälligen zeitvariierenden Parametern. Diploma thesis. ETH Zürich, Zürich.

Dal Molin, M., Schirmer, M., Zappa, M., Fenicia, F., 2020. Understanding dominant controls on streamflow spatial variability to set up a semi-distributed hydrological model: the case study of the Thur catchment. Hydrol. Earth Syst. Sci. 24 (3), 1319–1345.

Del Giudice, D., Albert, C., Rieckermann, J., Reichert, P., 2016. Describing catchment-averaged precipitation as a stochastic process improves parameter and input estimation. Water Resour. Res. 52, 3162–3186.

Doucet, A., Johansen, A.M., 2009. A tutorial on particle filtering and smoothing: Fifteen years later. Handbook of nonlinear filtering 12 (656–704), 3.

Doucet, A., Johansen, A.M., 2012. A tutorial on Particle filtering and smoothing. Handb, Nonlinear Filter, p. 12.

Evensen, G., 2009. Data assimilation: the ensemble Kalman filter. Springer Science & Business Media.

Evin, G., Kavetski, D., Thyer, M., Kuczera, G., 2013. Pitfalls and improvements in the joint inference of heteroscedasticity and autocorrelation in hydrological model calibration. Water Resour. Res. 49 (7), 4518–4524.

Evin, G., Thyer, M., Kavetski, D., McInemey, D., Kuczera, G., 2014. Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity. Water Resour. Res. 50, 2350–2375.

Fearnhead, P., Künsch, H.R., 2018. Particle filters and data assimilation. Annu. Rev. Stat. Its Appl. 5, 421–449.

Fenicia, F., Kavetski, D., Reichert, P., Albert, C., 2018. Signature-domain calibration of hydrological models using approximate bayesian computation: Empirical analysis of fundamental properties. Water Resour. Res. 54 (6), 3958–3987.

Fenicia, F., Kavetski, D., Savenije, H.H.G., 2011. Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development. Water Resour. Res. 47 (11).

Fenicia, F., Kavetski, D., Savenije, H.H.G., Clark, M.P., Schoups, G., Pfister, L., Freer, J., 2013. Catchment properties function, and conceptual model representation: is there a correspondence? Hydrol. Process. 28 (4), 2451–2467.

Foreman-Mackey, D., Hogg, D.W., Lang, D., Goodman, J., 2013. emcee: The MCMC Hammer. Publ. Astron. Soc. Pac. 125 (925), 306–312.

Jakeman, A., Hornberger, G., 1993. How much complexity is warranted in a rainfall-runoff model? Water Resour. Res. 29 (8), 2637–2649.

Kuczera, G., Kavetski, D., Franks, S., Thyer, M., 2006. Towards a bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters. J. Hydrol. 331 (1–2), 161–177.

Leisenring, M., Moradkhani, H., 2011. Snow water equivalent prediction using bayesian data assimilation methods. Stoch. Env. Res. Risk Assess. 25 (2), 253–270.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., Bergström, S., 1997. Development and test of the distributed hbv-96 hydrological model. J. Hydrol. 201 (1–4), 272–288.

Liu, Y., Gupta, H.V., 2007. Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework. Water Resour. Res. 43 (7).

Mandelbrot, B.B., Wallis, J.R., 1968. Noah, joseph, and operational hydrology. Water Resour. Res. 4 (5), 909–918.

McInerney, D., Thyer, M., Kavetski, D., Lerat, J., Kuczera, G., 2017. Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors. Water Resour. Res. 53 (3), 2199–2239.

McInerney, D., Thyer, M., Kavetsky, D., Lerat, J., Kuczera, G., 2017. Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors. Water Resour. Res. 53, 2199–2239.

MeteoSwiss, 2018. Meteoswiss: https://www.meteoschweiz.admin.ch/home/service-und-publikationen/beratung-und-service/datenportal-fuer-experten.html (last access: 19 June 2022), 2018.

Moradkhani, H., Hsu, K.-L., Gupta, H., Sorooshian, S., 2005. Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter. Water Resour. Res. 41 (5).

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part i–a discussion of principles. J. Hydrol. 10 (3), 282–290.

Perrin, C., Michel, C., Andréassian, V., 2003. Improvement of a parsimonious model for streamflow simulation. J. Hydrol. 279 (1–4), 275–289.

R Core Team, 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Reichert, P., Ammann, L., Fenicia, F., 2021. Potential and challenges of investigating intrinsic uncertainty of hydrological models with stochastic, time-dependent parameters. Water Resour. Res. 53 (3) e2020WR028400.

Reichert, P., Mieleitner, J., 2009. Analyzing input and structural uncertainty of nonlinear dynamic models with stochastic time-dependent parameters. Water Resour. Res. 45 (10).

Reichert, P., Schuwirth, N., 2012. Linking statistical bias description to multiobjective model calibration. Water Resour. Res. 48 (9).

Schirmer, M., Luster, J., Linde, N., Perona, P., Mitchell, E.A.D., Barry, D.A., Hollender, J., Cirpka, O.A., Schneider, P., Vogt, T., Radny, D., Durisch-Kaiser, E., 2014. Morphological hydrological, biogeochemical and ecological changes and challenges in river restoration – the Thur River case study. Hydrol. Earth Syst. Sci. 18 (6), 2449–2462.

Schoups, G., Vrugt, J.A., 2010. A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-gaussian errors. Water Resour. Res. 46 (10).

Sikorska, A.E., Renard, B., 2017. Calibrating a hydrological model in stage space to account for rating curve uncertainties: general framework and key challenges. Adv. Water Resour. 105, 51–66.

Tomassini, L., Reichert, P., Künsch, H.R., Buser, C., Knutti, R., Borsuk, M.E., 2009. A smoothing algorithm for estimating stochastic, continuous-time model parameters and its application to a simple climate model. J.R. Statist. Soc. C: Appl. Stat. 58, 679–704.

Uhlenbeck, G.E., Ornstein, L.S., 1930. On the theory of the Brownian motion. Phys. Rev. 36 (5), 823.

van Leeuwen, P., Künsch, H., Nerger, L., Potthast, R., and Reich, S. (2019). Particle filters for high-dimensional geoscience applications: A review. Q J R Meteorol Soc, 145: 2335–2365.

Van Leeuwen, P.J., Künsch, H.R., Nerger, L., Potthast, R., Reich, S., 2019. Particle filters for high-dimensional geoscience applications: A review. Q. J. R. Meteorol. Soc. 145 (723), 2335–2365.

Verleysen, M., François, D., 2005. The curse of dimensionality in data mining and time series prediction. In: International work-conference on artificial neural networks. Springer, pp. 758–770.

Vrugt, J.A., ter Braak, C.J., Diks, C.G., Schoups, G., 2013. Hydrologic data assimilation using particle Markov chain Monte Carlo simulation: Theory concepts and applications. Adv. Water Resour. 51, 457–478.

Wagener, T., Boyle, D.P., Lees, M.J., Wheater, H.S., Gupta, H.V., Sorooshian, S., 2001. A framework for development and application of hydrological models. Hydrol. Earth Syst. Sci. 5 (1), 13–26.

Wagener, T., McIntyre, N., Lees, M., Wheater, H., Gupta, H., 2003. Towards reduced uncertainty in conceptual rainfall-runoff modelling: Dynamic identifiability analysis. Hydrol. Process. 17 (2), 455–476.

Šukys, J., Bacci, M., 2021. Spux framework: a scalable package for bayesian uncertainty quantification and propagation.