Exploring a copula-based alternative to additive error models—for non-negative and autocorrelated time series in hydrology

June 27, 2019

Omar Wani[1,2], Andreas Scheidegger[1], Francesca Cecinati[3],Gabriel Espadas[1,4], Jörg Rieckermann[1]

[1] Eawag, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland

[2] Institute of Environmental Engineering, ETH Zürich, Zürich, Switzerland

[3] Department of Architecture and Civil Engineering, University of Bath, Bath, UK

[4] (formerly at) Seminar for Statistics, ETH Zürich, Zürich, Switzerland

Published article: Wani, O., Scheidegger, A., Cecinati, F., Espadas, G., Rieckermann, J. (2019). Exploring a copula-based alternative to additive error models—for non-negative and autocorrelated time series in hydrology. Journal of Hydrology, 575(August 2018), 1031–1040. https://doi.org/10.1016/j.jhydrol.2019.06.006

Abstract

Inaccurate description of uncertainty in the error models can cause biases in parameter estimation. When the parameters of the deterministic model and the error model are inferred jointly from the observations, the posterior converges to regions that reflect the processes in both high and low flows. If the nature of errors in low and high flows is different to the extent that the same error description cannot be used for both, biases in inference are introduced. In such cases, the parameter posterior will adjust to the region of the hydrograph with longer proportionate presence in the calibration time series. In this paper we demonstrate that the autoregressive order 1 (AR1) description of errors can lead to sub-optimally performing predictive models if the calibration period has substantial sections of inadequately modelled flows. Inference is performed within the Bayesian framework. We show this for a synthetic example as well as a case study. We also see that the predictive uncertainty bands that we get using the AR1 description can be overconfident and also admit negative values. To mitigate this, we analyze an alternative to additive error models. We use a distribution with a non-negative support, gamma in this study, reflecting the uncertainty in the system response at every time step. The gamma distribution is conditioned on the deterministic model output, which determines its mode and standard deviation. We capture autocorrelation in time using copulas. Given that copulas can capture dependence between different marginals, we use different specifications of the marginal distribution for high and low flows. The results show that 1) biases in parameter estimation can be reduced if a representative error description is attained using the flexibility of a copula-based likelihood. 2) The non-negative support allows to make more realistic uncertainty intervals for low flows. 3) However, the autocorrelation parameter in copulas severely interacts with the model and heteroscedasticity parameters. 4) While the formulation, in principle, should be of added value for parameter inference, in case of less informative priors, the flexibility of this description can produce non-robust inference.

1 Introduction

Due to the complexity of hydrologic processes like rainfall-runoff, deterministic modelling proves to be inadequate for efficient decision making (Krzysztofowicz, 2001; Verkade and Werner, 2011). Use of probability theory helps quantify uncertainty in such predictions (Gupta et al., 1998; Honti et al., 2013; Kuczera et al., 2006; Refsgaard et al., 2007). Specifically within the Bayesian framework, to further constrain the range of parameters values that seem feasible for a catchment a prioiri, observations of the system response are used for parameter probability density updating. The physical understanding of the system, as well as the understanding of errors is put in formulating the conditional probability density of system response, $p(\boldsymbol{y_o} \mid \boldsymbol{x}, \boldsymbol{\theta})$, given parameters ($\boldsymbol{\theta}$) and input forcing ($\boldsymbol{x}$). This reflects our assumption about the system as the observation generating random processes. Bayes' theorem is then used to invert the conditioning and get the probability density of parameters given some observations of the system i.e. $p(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{y_o})$ (Hall et al., 2011; Kavetski, 2018; Kennedy and O'Hagan, 2001).

However, like the assumptions about the deterministic model, our assumptions about the errors should be representative to an adequate degree so that inference gives meaningful parameter estimates. The rainfall-runoff time series usually has these properties: 1) scaling with rainfall, 2) autocorrelation in time, and 3) heteroscedasticity i.e. changing variance and 4) non-negativity. While the deterministic rainfall-runoff models provide information about the aggregate tendencies of the flow, for an unbiased parameter estimation and accurate uncertainty estimation, the hydrologic error model should be able to reproduce the other stochastic properties to an adequate degree (Del Giudice et al., 2016). However, it is not straightforward to reproduce all of these properties in the formulation of the likelihood function without some further simplifying assumptions, for example allowing the possibility of negative flow predictions in additive error models (Sun et al., 2017). Also, the structural deficits in many hydrologic models may come from very systematic biases which may not lend themselves adequately to one single probabilistic description for the entire hydrologic time series. Data from different hydrologic regimes, like low and high flows, may appear to be a result of different data generating random processes that need their own unique probabilistic model, e.g different autocorrelation (Ammann et al., 2019).

Poor accounting of low flow errors can otherwise lead to poor prediction for high and medium flows. Generally the modeler prioritizes one property over the other, depending on the task for which the model is being designed.

To make error descriptions more representative of the rainfall-runoff, several parameterizations for the likelihood function have been proposed, like for example, to capture heteroscedasticity (Evin et al., 2013) , skewness (Schoups and Vrugt, 2010), autocorrelated structural deficits through time continuous processes (Del Giudice et al., 2013) etc. Beyond using probabilistic description of errors to perform parameter inference for the model, these descriptions have also been utilized to just quantify predictive uncertainty, given a hydrologic model with fixed parameters (Weerts et al., 2011). Flexibility in the description of errors has been sought for a more representative parameter estimation and a more reliable predictive uncertainty estimation, be it within a formal Bayesian framework or as a post-processor where parametric uncertainties are neglected e.g (Dogulu et al., 2015; Pianosi and Raso, 2012; Wani et al., 2017). However, especially for inference, it has been challenging to suggest a probabilistic description of errors that is effective for all different catchments and all different models (McInerney et al., 2017).

One of statistical tools to capture more complex multivariate probability densities in hydrology, be it data or modelling, is copulas (Bárdossy and Hörning, 2016; Sadegh et al., 2017; Salvadori and De Michele, 2004). Especially in the context of uncertainty analysis, copulas have been used to model the residual uncertainty in the post-processing phase of the model (Klein et al., 2016; Liu et al., 2018). The ability of copulas to model dependence between random variables, regardless of the nature of their marginal distributions, gives us the choice to construct the description of marginals and dependence separately. One of the limitations of exploring more representative marginals of errors to construct likelihood functions is the loss of temporal autocorrelation. Copulas can then be used to model such time series dependence e.g (Borgomeo et al., 2015; Salvadori and De Michele, 2004).

In this paper, we describe and analyze a relatively general formulation of likelihood function for hydrologic models. This formulation employs copulas, which offer much more flexibility, for example by allowing the parameterization of errors separately for low and high flow, while having the possibility to represent autocorrelation and non-negativity. The additive AR1 Gaussian error model can be seen as a special case of this general formulation. Using results from simulation experiments, we move on to show the effects of inadequately modeled low flows on parameter inference. We do this using synthetic data and real data. We also show that the flexibility in the proposed formulation allows the modeler to choose different parameterized marginals for high and low flows and we discuss the benefits and shortcoming of this formulation.

2 Method

2.1 Error Model

Additive description: If we expect to observe a hydrologic system response $\boldsymbol{Y_o} = [Y_{o,1}, \ldots, Y_{o,T}]$, given input time series $\boldsymbol{x}$ and a model output $\boldsymbol{y_m} = [y_{m,1}, \ldots, y_{m,T}]$ with parameters $\boldsymbol{\theta}$, the classical description of additive autoregressive Gaussian error $\boldsymbol{B}$ gives us (vectors in bold and random variables in capital letters):

$$\boldsymbol{Y_o}(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\psi}) = \boldsymbol{y_m}(\boldsymbol{x}, \boldsymbol{\theta}) + \boldsymbol{B}(\boldsymbol{x}, \boldsymbol{\psi}) \tag{1}$$

where $\boldsymbol{\psi}$ is the error model parameter vector. It follows that the conditional probability density for T observations , $\boldsymbol{y_o} = [y_1, ..., y_T]$, given a certain input and parameter vector, is a multivariate normal distribution

$$p(\boldsymbol{y_o} \mid \boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\psi}) = \mathcal{N}\big(\boldsymbol{y_o} \mid \boldsymbol{y_m}(\boldsymbol{x}, \boldsymbol{\theta}), \Sigma(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\psi})\big) \tag{2}$$

with the mean given by the hydrologic model and the covariance matrix Σ is derived from a covariance function that models autocorrelation. Samples from this probability density give us “candidate” time series of observations. We expect the observation to be *like* those samples. The whole range of the samples constitutes the predictive distribution of the hydrologic system response, for a certain value of inputs and parameters. Then suitable quantiles from this distribution can be chosen to provide uncertainty bands.

One of the shortcomings is that a Gaussian process cannot guarantee non-negative model realizations. It would be desirable to use distributions with positive support as marginal distributions for each time step, such as gamma or lognormal distributions. However, this makes the incorporation of the autocorrelation structure difficult. Ammann et al., 2019, address this problem by transforming the real observation space, which does not admit negative values, to a space in which the distribution is expected to be normally distributed and then they model the autocorrelation using the standard Gaussian process.

Copula-based description: In this study, we present a generalized formulation to model autocorrelated time series. First, we select a familiy of distributions for the marginals, i.e. the system response at single point in time, and then use a copula that captures autocorrelation. The marginal distribution $f_t(y_{o,t})$ at time t of system response $Y_{o,t}$ should depend on the model output. The deterministic model $y_{m,t}$ describes the location (e.g. mean or mode) of f_t. The variance is also linked to the model using some linear or non-linear relationship, such that $\mathrm{sd}[Y_{o,t}] = g(y_{m,t}, \phi)$. The selection of the distribution family is guided by system understanding, for example negative support and multi-modal distributions should be avoided. To capture the temporal correlation between $f_1 \dots f_T$, a copula is introduced.

A T-dimensional copula is a cumulative probability distribution function defined over a unit cube in some T-dimensional space such that the marginal density over each dimension is uniform. Using a copula with density $c(u_1, ..., u_T)$ and the marginals, any joint distribution of the observations can be expressed as:

$$p(\boldsymbol{y}_o \mid \boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\psi}) = c\Big(F_1\big(y_{o,1} \mid y_{m,1}(\boldsymbol{x}, \boldsymbol{\theta}), \boldsymbol{\psi}\big), ..., F_T\big(y_{o,T} \mid y_{m,T}(\boldsymbol{x}, \boldsymbol{\theta}), \boldsymbol{\psi}\big)\Big) \prod_{t=1}^{T} f_t(y_{o,t} \mid y_{m,t}(\boldsymbol{x}, \boldsymbol{\theta}), \boldsymbol{\psi}) \tag{3}$$

where $F_t()$ is the cumulative probability distribution of the corresponding density $f_t()$ (Loaiza-Maya et al., 2018). The copula can be simplified if a Markov property is assumed for the time series. For example with an order-1 Markov property we can write Eq. (3) with a 2-dimensional copula

$$p(\boldsymbol{y}_o \mid \boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\psi}) = \prod_{t=2}^{T} c\Big(F_{t-1}, F_t\Big) \prod_{t=1}^{T} f_t \tag{4}$$

(with conditionals dropped for clarity). If the marginals are taken as normally distributed and a Gaussian copula is used to model the autocorrelation, this equation can be further reduced to an AR1 process.

Specifications for this study: In this study Frank copula, which is symmetric for low and high quantiles, is used to capture autocorrelation. Archimedean copulas, like Frank, have few parameters, and therefore do not increase the dimensionality of parameter space by much. Frank copula has one parameter ψ_α for controlling correlation and it is defined as:

$$C(u_1, u_2) = -\frac{1}{\psi_\alpha} \ln\Big[1 + \frac{(e^{-\psi_\alpha u_1} - 1)(e^{-\psi_\alpha u_2} - 1)}{(e^{-\psi_\alpha} - 1)}\Big] \tag{5}$$

Its density function is defined as:

$$c(u_1, u_2) = \psi_\alpha \frac{(1 - e^{-\psi_\alpha})(1 - e^{-\psi_\alpha (u_1 + u_2)})}{((1 - e^{-\psi_\alpha}) - (1 - e^{-\psi_\alpha u_1})(1 - e^{-\psi_\alpha u_2}))^2} \tag{6}$$

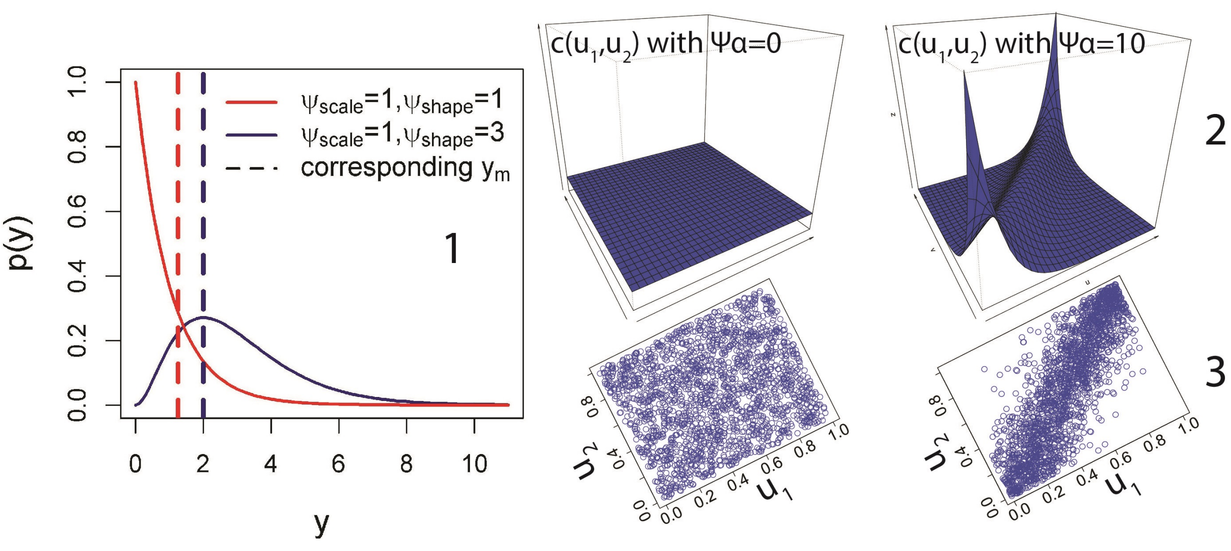

Figure 1: (1) Univariate gamma distribution with different ψ_{shape} and ψ_{scale} parameter values. (2) Density of Frank copula with different ψ_α values. (3) Corresponding samples.

To assure non-negative outputs $\boldsymbol{y_o}$ we use gamma distributed marginals $f_t()$:

$$Y_{o,t} \sim f_t(y \mid \psi_{scale}, \psi_{shape}) = \frac{1}{\psi_{scale}^{\psi_{shape}} \Gamma(\psi_{shape})} y^{(\psi_{shape}-1)} e^{-(\frac{y}{\psi_{scale}})} \tag{7}$$

where $\Gamma()$ is the gamma function.

We assume a threshold parameter, that separates low flows and high flows, called ψ_{base}. For low flows, we assume that the model has deficits that cause it to have errors such that it follows exponential distributions. So for this regime of flow we assign a shape factor of 1 and get the scale from the model output. For flows higher than ψ_{base}, the assumption is made that the deterministic model output is the mode of the observational distribution (Fig. 1). This differentiates the observation generating process for low and other flows.

In this research, the standard deviation (σ_t) of $f_t()$ is as a function of the deterministic model output. It is incorporated by the explicit equivalent of Box-Cox, with $\lambda = 0.5$

$$\mathrm{sd}(Y_{o,t}) = \sigma_t = g(y_{m,t}, \psi_1, \psi_2) = \psi_2(y_{m,t} + \psi_1)^{1-\lambda} \tag{8}$$

If we would have a normal distributed additive error, this scaling would correspond to a Box-Cox transformation. This result can be derived by using the truncated Taylor approximation (McInerney et al., 2017). To illustrate this dependency we sampled an additive normal distributed error term and applied the inverse Box-Cox transformation (see Fig. 2).

Defining the standard deviation dependence explicitly makes the formulation of priors for the error model parameters more intuitive, compared to the traditional Box-Cox transformation. Here ψ_1 and ψ_2 are parameters to be inferred . All this defines the parameters of the gamma distribution as:

$$Y_{o,t} \sim \begin{cases} f_t\big(y|\psi_{shape} = 1, \psi_{scale} = \sigma_t\big) & if\ y_{m,t} < \psi_{base} \\ f_t\big(y|\psi_{shape} = \big(\frac{y_{m,t}}{\psi_{scale}} + 1\big), \psi_{scale} = \big(\frac{\sqrt{4\sigma_t^2 + y_{m,t}^2} - y_{m,t}}{2}\big)\big) & \text{otherwise} \end{cases} \tag{9}$$

This definition of ψ_{shape} and ψ_{scale} for $y_m \geq \psi_{base}$ assures that the mode of the distribution is $y_{m,t}$ and its standard deviation is σ_t.

Given the general formulation of Eq. (4), we are free to pick and choose different autocorrelation structures (similar to Eq. (6)), heteroscedasticity structures (similar to Eq. (8)) and marginal distributions (similar to Eq. (9)). Each of these three components can be employed independently of the other, with its own parameterization.

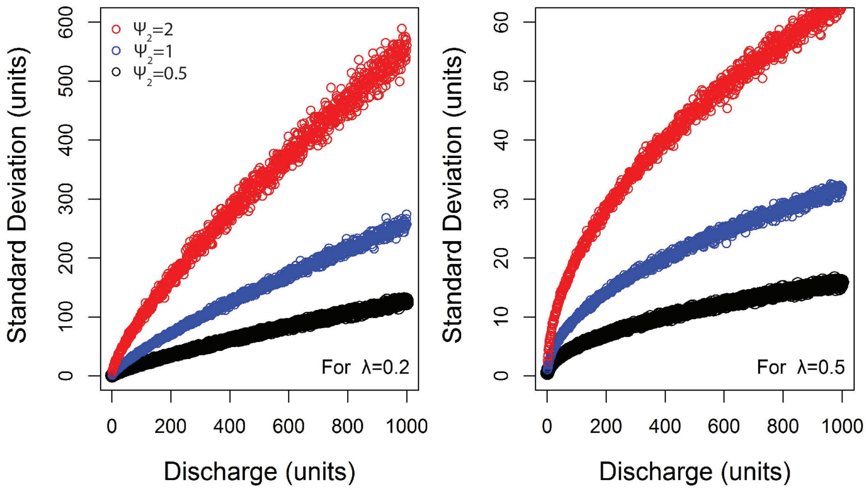

Figure 2: The change in standard deviation (σ_t) of the error in untransformed space, when the error is assumed to be homoscedastic (with a fixed standard deviation $= \psi_2$) in the Box-Cox transformed space. This graph helps to understand the implicit dependence of standard deviation of the output when using Box-Cox. Using Taylor approximation this dependence is $\sigma_t = \psi_2 (y_m)^{(1-\lambda)}$.

2.2 Bayesian inference

The expressions in Eqs. (2) and (4) turn into a likelihood function when a given set of observations $\boldsymbol{y_o}$ are put into them and the parameter values are varied. This expression is not a probability density in the parameter space $(\boldsymbol{\theta}, \boldsymbol{\psi})$, but only in the observation space $(\boldsymbol{y_o})$, hence referred to as a function of $(\boldsymbol{\theta}, \boldsymbol{\psi})$. As likelihood function gets defined uniquely once we define the probability model for Y_o, therefore, in this paper 'copula-based probability model' and 'copula-based likelihood function' refer to the same mathematical description as defined in section 2.1. Once the likelihood function is specified, in combination with the prior, we use Bayes' theorem to get the posterior. An adaptive Markov Chain Monte Carlo (MCMC) sampling scheme is employed to get the parameter samples (Vihola, 2012; Scheidegger, 2018). The samples represent the updated belief in the parameter values, given the observed data.

$$p(\boldsymbol{\theta}, \boldsymbol{\psi} | \boldsymbol{y_o}, \boldsymbol{x}) = \frac{p(\boldsymbol{y_o} | \boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\psi}) p(\boldsymbol{\theta}, \boldsymbol{\psi})}{\int p(\boldsymbol{y_o} | \boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\psi}) p(\boldsymbol{\theta}, \boldsymbol{\psi}) d\boldsymbol{\theta} d\boldsymbol{\psi}} \tag{10}$$

These samples are then run through the probabilistic model to generate the prediction bands.

2.3 Rainfall-runoff model

For analysis using synthetic data, we use a linear model and with input P and (θ_1, θ_2) as model parameters.

$$y_{m,t} = \theta_1 P_t + \theta_2 \tag{11}$$

This model gives flexibility and speed to facilitate preliminary analysis and provide proof of concept.

To analyze the ability of this likelihood formulation on real data, we use a simple conceptual model with deficits which do not take infiltration, evapotranspiration and rainfall variability into account. A unit hydrograph convolution is used which defines the input-output relationship between discharge and precipitation. Such models have been used before for capturing discharge relationship of catchments (Betterle et al., 2017). The model assumes an exponentially decaying response to a unit of instantaneous rainfall. Given this formulation, the response of a time series of rainfall, $P(t)$, can be obtained by integrating the response corresponding to each time slice of the rainfall .

$$y_{m,t} = \theta_A \int_0^t P(t - \tau) \cdot \theta_k e^{-\theta_k \tau} d\tau \tag{12}$$

Parameter θ_A represents the effective area of the catchment, and θ_k represents the dependence on the rainfall at past time steps. The motivation here is to use fast and simple rainfall-runoff models with overt deficits so that the performance of the error description can be evaluated.

2.4 Case study

The study area is the catchment of the River Rawthey, North of England, in the Yorkshire Dale National Park, an affluent of the River Lune. It covers 219 km^2, and collects waters from a terrain that ranges from 675m to 85m of elevation. The catchment is primarily a Natural Park, with very limited human intervention in terms of agriculture and urban land use. The rainfall data used in

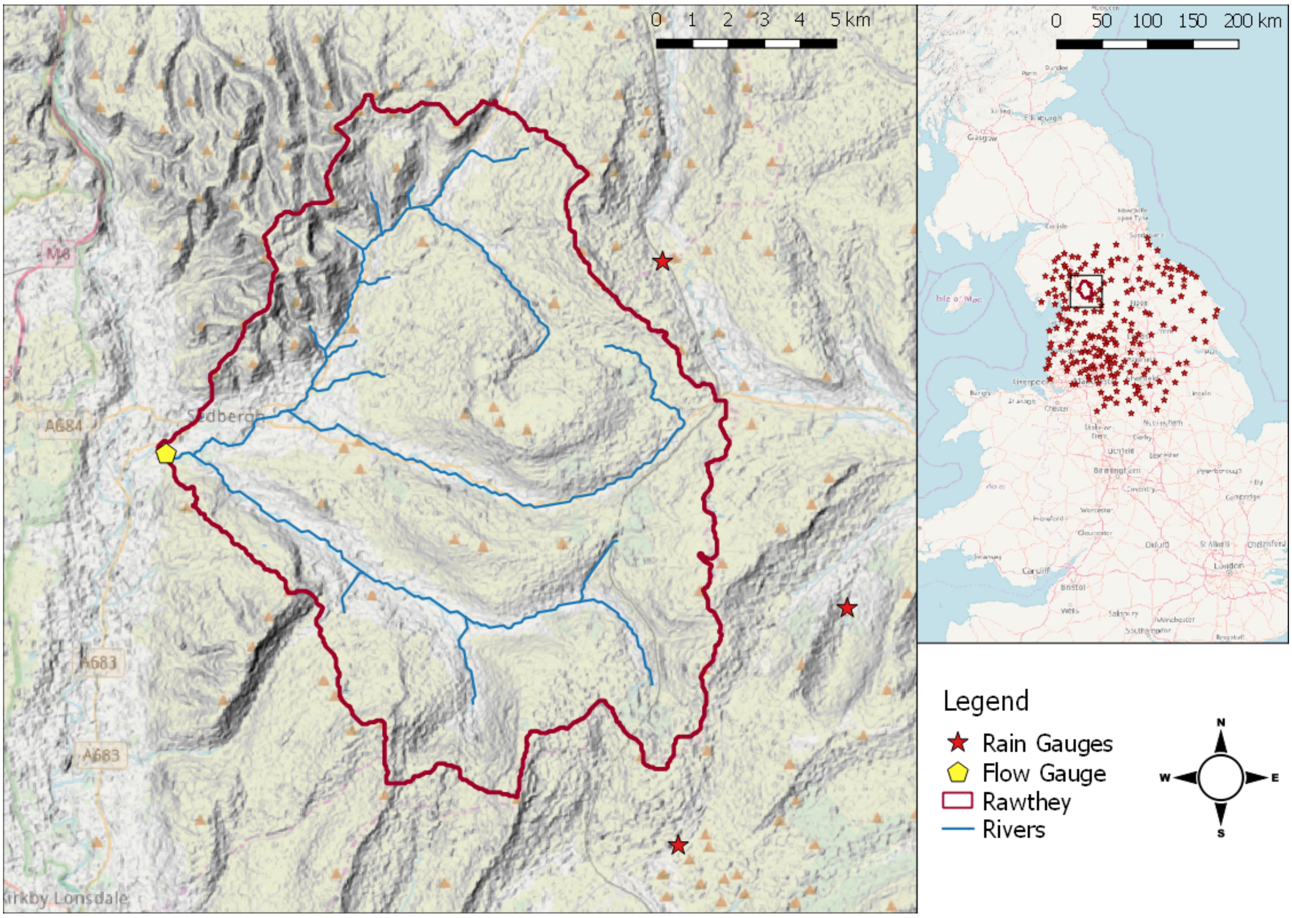

Figure 3: The Rawthey catchment in England.

this work is obtained from rain gauges from the Environment Office. Data was available from a set of 203 tipping bucket devices with a resolution of 0.2 mm, available at 15 minutes resolution, then accumulated at hourly resolution for two years, from January 2009 to December 2010. The study area, the rain gauges, and the flow gauge are represented in Fig. 3. Rainfall time series are obtained with block kriging (Chiles and Delfiner, 2009). Block kriging is a technique that integrates the values obtained through ordinary kriging over an area. Apart from this, we also use a rainfall multiplier, named θ_r, as a parameter for inference. This is done to alleviate the systematic errors due to the fact that none of the rain gauges used to generate the rainfall time series lie within the catchment area.

2.5 Simulation experiments

Parameter estimation for synthetic data:

A) We first show using a didactic example why incorporating autocorrelation and heteroscedasticity is important for unbiased parameter estimation and predictive uncertainty estimation. We do this using synthetic data generated from a linear model (Eq. (11)), feeding it with a sinusoidal rainfall. The error is assumed to be heteroscedastic, following Eq. (8), and autocorrelated (AR1), with a correlation parameter ψ_ρ. Inference is done using flat priors. The parameter values for $(\theta_1, \theta_2, \psi_1, \psi_2, \psi_\rho)$ used to produce data are (1,5,2,2.5,0.9). ψ_1 is kept fixed. Inference is done using $\boldsymbol{P}$ as, $30\sin(\boldsymbol{x}) + 30$, with $\boldsymbol{x}$ from 0 to 100, spaced at 0.1. Verification time series is generated as a second sample using same probability model specification.

B) Preliminary proof of concept: To facilitate reproducibility, we again do parameter inference using the new copula-based formulation of likelihood and synthetic data generated through same linear model. Inference is done using flat priors. Eqs. (8) and (11) are used to generate synthetic data. We use two precipitation time series:

$$\boldsymbol{P_{high}} = 30\sin(\boldsymbol{x}) + 30, \quad \boldsymbol{P_{low}} = 10\sin(\boldsymbol{x}) + 10 \tag{13}$$

A sequence of $\boldsymbol{x}$ from 0 to 100, spaced at 0.1, is used to generate P at discreet time steps. We then combine first 220 points from $\boldsymbol{P_{low}}$ and the other 780 from $\boldsymbol{P_{high}}$ to get the precipitation for the calibration time series. And combine first 420 points from $\boldsymbol{P_{low}}$ and the other 580 from $\boldsymbol{P_{high}}$ to get the precipitation for the validation time series. The parameters (θ_1, θ_2) used for the high flow are (1,5) and for the low flow (0.1,5). This gives us observation generated by two separate processes. ψ_2 and ψ_ρ are 0.3 and 0.9 respectively.

<u>Parameter estimation for real case study</u>: For the case study we use two time series of observation for calibration and validation. As there are substantial sections of the time series that are low flows, we see the effect of such flows on the inference of the parameters and on the predictive uncertainty. We use an hourly time series of discharge and precipitation. We calibrate the model on 2000 data points, from 01.01.2009 00:00 to 25.03.2009 07:00, and then validate the results for 3000 data points, from 08.09.2009 00:00 to 10.01.2010 23:00. Both the calibration and validation time series contain sections of high and low flows. As the model used is in this study has substantial deficits, we expect the prediction intervals to be wide. The model has two inference parameters θ_A and θ_k (Eq. (12)).

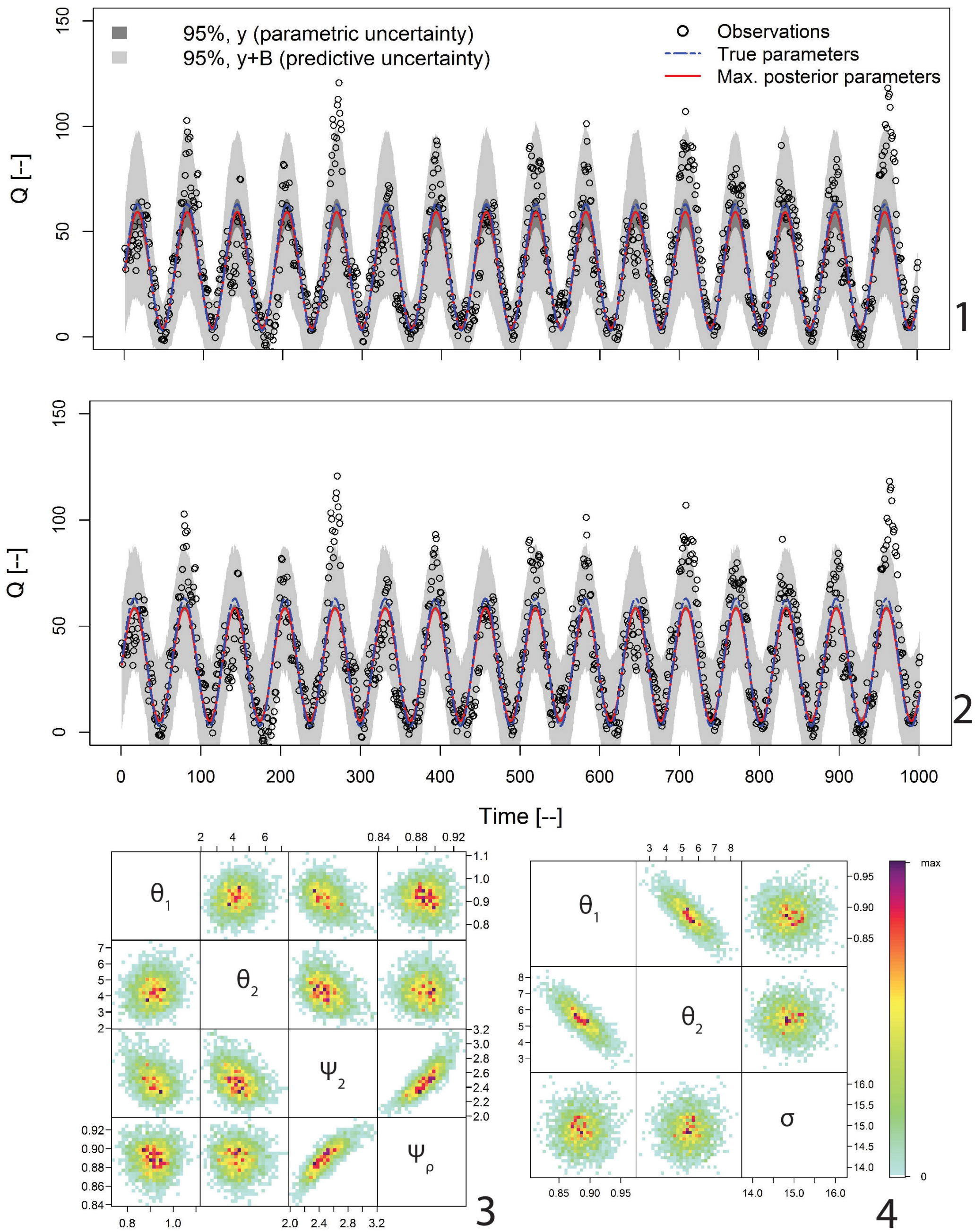

Figure 4: Effect of various assumptions regarding the likelihood on the inference of linear model parameters when the error generating process is alike for low and high flows (synthetic data from an additive error model). 1) Predictive uncertainty estimated using correlated and heteroscedastic (true) likelihood 2) Predictive uncertainty estimated using uncorrelated and homoscedastic (i.i.d) likelihood, with standard deviation σ. 3) and 4) are the corresponding posteriors.

The priors used for the inference are mentioned are presented in Table 1. And the ratio high and low flows in the calibration and validation time series is presented in Table 2.

Table 1: Normal truncated distributions were used to define the priors for the parameters in simulation experiments for the case study. The vector in the table give the values for mean, standard deviation, lower limit and upper limit of the prior for each parameter in our inference.

For copula-based likelihood function						
ψ_1	ψ_2	ψ_α	θ_A	θ_r	θ_k	ψ_{base}
(5,2,0,30)	(0.4,0.3,0,5)	(5,10 ,0,50)	(16000,10000, 12000,100000)	(1.5,1, 1,5)	(0.2,0.5, 0.08,3)	(50,3, 40,60)

For AR1 likelihood function					
ψ_1	ψ_2	ψ_ρ	θ_A	θ_r	θ_k
(5,2,0,30)	(0.4,0.3,0,5)	(0.9,0.3,0,1)	(16000,10000, 12000,100000)	(1.5,1, 1,5)	(0.2,0.5, 0.08,3)

Table 2: High and low flows in calibration and validation time series.

Calibration			Validation		
Highest flow (m^3/s)	No. of observations below 30 m^3/s	Number of observations above 150 m^3/s	Highest flow (m^3/s)	No. of observations below 30 m^3/s	No. of observations above 150 m^3/s
259	1901	12	279	2607	14

3 Results

As it is evident from Fig. 4, when data that is generated from an autocorrelated and heteroscedastic process is used to infer parameters for a linear model, the inference results in underestimation of parametric (given by the spread in the posterior) and predictive uncertainties (given by the width of prediction intervals for high flows). Also, in the predictive phase the high and low flows do not get captured adequately. For example in this case, the prediction intervals using independent, identically distributed (i.i.d) model are narrower for high flows and wider for lower flows, compared to actual distribution of observation in these flow regimes. Also, the posteriors of parameters from i.i.d are narrower than the posterior from autocorrelated model. We additionally see that inference under the i.i.d assumption can show artificial dependence between various parameters, e.g. high correlation between (θ_1, θ_2), which is not the case when using the true error model. This didactic example establishes the operational need to capture the autocorrelation and heteroscedasticity of a hydrologic time series.

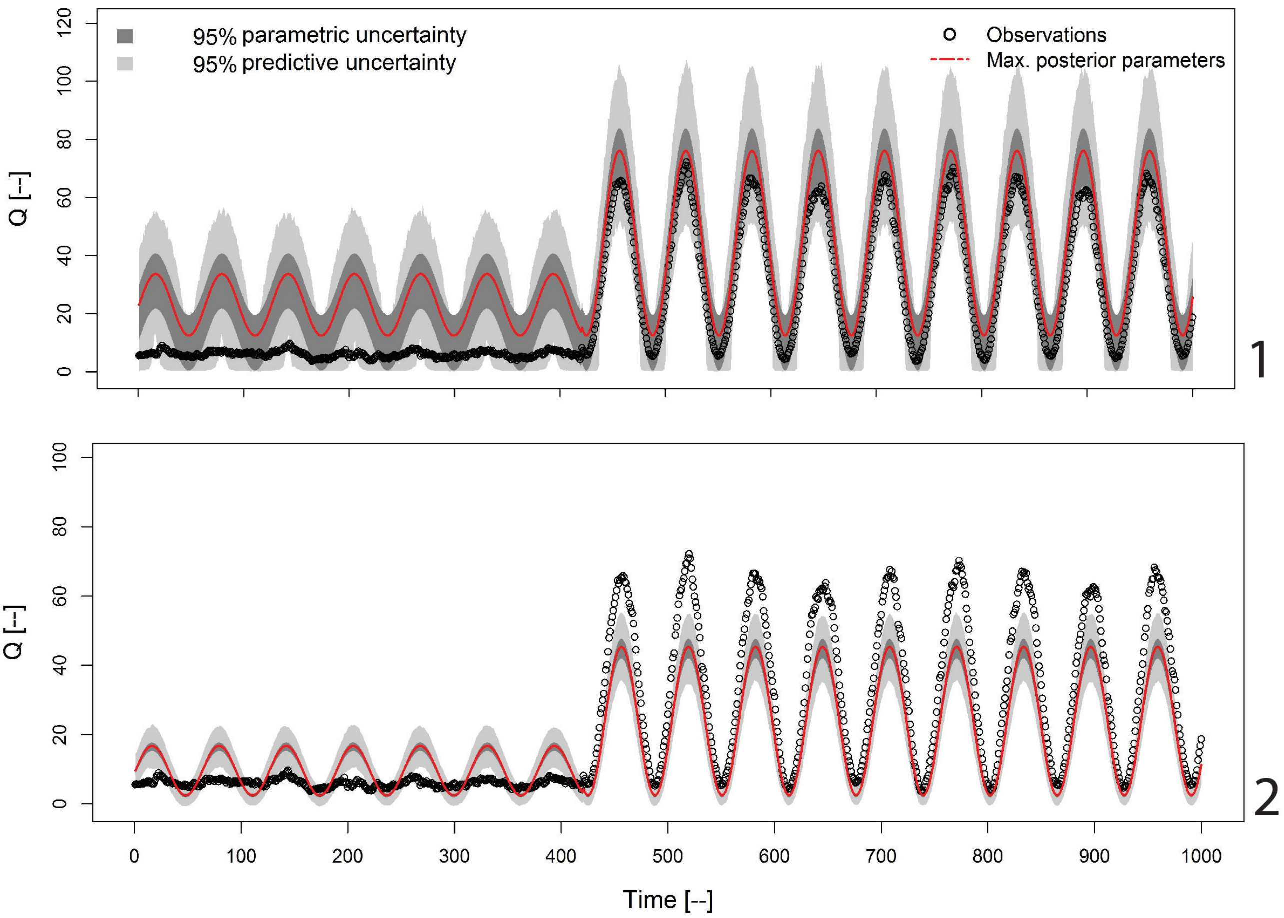

Figure 5: Model deficits in low flows inducing biased parameter inference can be avoided by using the flexibility of copula based likelihood during inference (synthetic data). 1) Prediction using the copula based probability model. 2) using AR1 probability model.

Apart from heteroscedasticity and autocorrelation in the hydrologic time series, errors in high and low flows can be generated by different deficits in the model, and therefore it may not be reasonable to treat them with the same marginal error distribution. Fig. 5 depicts the inference of the linear model which systematically overestimates low flows. This bias in the model causes severe biases in inference if we use the AR1 processes. The model parameters tend to underestimate flows during high rainfall, so that they can fit the base flows better. One way to overcome this problem is using a different type of marginal distribution, which can capture such a tendency of the modelling errors. The error description, as suggested in Eq. (9), is more tolerant towards the biases in the low flows and thus allows the model to fit for high flows Also, we see that the non-negativity of the flow is avoided.

Similar underestimation of high flows is seen for the real case study. Whereas copulas-based likelihood description is able to save the inference procedure from the underestimation of high flows (Fig. 6). We use the maximum posterior density parameter values as well as the parameter values from the whole posterior density to assess the prediction intervals. Given we anticipate that model does not capture the infiltration process well, and may overestimate runoff from small rainfall events, we put this understanding in Eq. (9), where low flow predictions are more likely to overestimate flow. This makes the probability model lenient towards such errors in low flows, giving the inference procedure the flexibility to search for parameters that take the prediction closer towards the high flow observations. Many other formulations, different from Eq. (9), can be used to incorporate more complex error structures.

From the analysis of the real case study, we find that around 62% of observational time steps for the validation time series have prediction intervals going into negative flows when only using the maximum posterior parameter values; this number increases to 67.4% when using the whole posterior density. No such negative flows are seen using the copula-based likelihood function description. Also, for 95% prediction interval, the mean width of the copula-based likelihood is higher, and it thus captures more observations within it. AR1 captures 81.7% and copula-based probability model 92.1% observations in the validation phase, using the maximum posterior parameter values; using the whole posterior density the numbers go up to 86.8% and 96.2% respectively. For the high flow regime (flows higher that 150 m^3/s), the relative mean error of model prediction corresponding to maximum posterior parameter values is 33.2% for AR1 and 14.2% for copula-based likelihood function.

4 Discussion

The results of the simulation experiments demonstrate that, akin to the accuracy of the deterministic hydrologic model, our description of the errors should also be representative of the modelled process. A departure from this can lead to non-trivial biases in the parameter estimation (Fig. 4 and 5). We also show that the disproportionate presence of low and high flows affects how the parameter inference would perform (Fig. 5 and 6). Manually choosing only high events from the past time series to calibrate the model is not always a desirable alternative. We may also be interested in the prediction of low flows using the same model. Then the error description needs to be flexible enough to capture both high and low flows adequately. There can be cases where the errors in the low flows are skewed, with the model having some systematic tendencies to depart from the observations. It is intuitive that the tendency of the model to overestimate and underestimate would not be symmetric for flows closer to zero. The errors have to truncate such that the real flow is always positive. In order to prevent such complex errors undesirably influencing the whole inference procedure, we can describe these errors with heavy-tail or exponential distributions (Eq. (9)). The results show that AR1, due to its inflexibility with the autocorrelation is not able to capture non-negative support and skewed distributions simultaneously. We use copulas to capture this and see a noticeable improvement in inference. Copulas allow us to capture the temporal dependence and we are free to choose the marginals. However, we also found that this flexibility, while in principle desirable, in practice still does not guarantee unbiased parameters. As we can see from Fig. 6 (3), ψ_α parameter tends to go towards the upper extreme value, and is only bound by the hard prior. While this is not the case for the synthetic case studies, where the error structures are known, real case studies have more complex errors and these interactions between error model parameters and deterministic model parameters become more severe. So this description of error, as defined in Eq. (4), just makes an unbiased inference potentially achievable, if we are able to define the adequately representative marginal distributions using this flexibility, and have informative priors, for example in the case of Fig. 6. More concrete guidelines to choose such representative marginals still need to be researched. And it is not possible to formulate a unique description of errors that is representative of all the hydrologic time series.

We find that copula-based likelihood functions prefer a high correlation value, as such parameter values produce peaky densities at the edges of the unit cube (Fig. 1). These edges correspond to high and low quantiles. In principle, high copula density values at these edges should be compensated by corresponding low values of marginal densities, hence avoiding parameter biases. As the likelihood is a product of copula densities and the marginal densities (Eq. (4)), if copula densities have high value

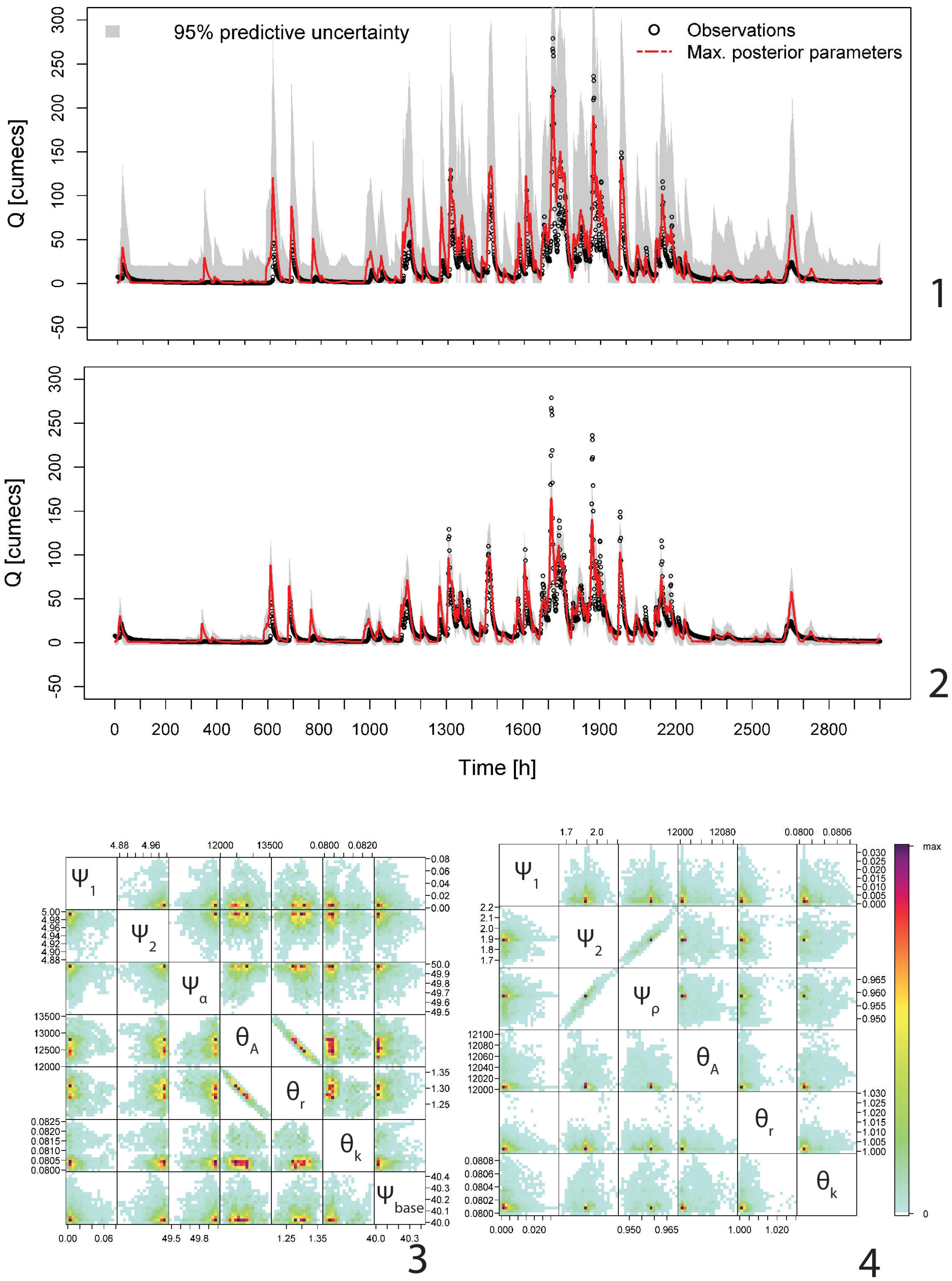

Figure 6: Comparing inference using copula based likelihood and AR 1 likelihood. 1) and 2) are the predictions. And 3) and 4) are the bivariate posteriors for copula based likelihood and AR1 respectively.

for extreme correlations, the low values of marginal should bring down the product and we should still have low likelihood values. However, this is not always achieved - the observations are not always described adequately by the selected probability model, so the compensation of the marginal is not guaranteed to suffice in avoiding very high parameters values of copulas. In many cases the posterior can converge to a high autocorrelation, at the expense of taking the deterministic model away from the observational data. One way to avoid this is by having informative priors where we know for sure that the autocorrelation of the process cannot be more than a certain amount and the standard deviation of the error cannot be more than a certain amount, representing the model understanding and the local hydrology of that particular case study. Nonetheless, more robust schemes which allow the model to go closer to the observations even when the priors are uninformative need to be formulated. The desirable formulation an error model would be such that small inaccuracies in the error models only lead to small inaccuracies in inference. Priors, however, can be used to exclude unreasonable parameter spaces.

As far as the improved performance of copula-based likelihood function over the AR1 is concerned, one of the reasons why this is seen in both synthetic and the real case studies of this work is that the peaky low flows are poorly represented by the chosen hydrologic models. The rainfall produces a flow signal in the model which is not seen in the real catchment. This effect is a combination of several sub-effects. First, the rainfall itself is measured by rain gauges that do not lie within the catchment (Fig. 3). The biases in the precipitation input, with overestimation and underestimation of rainfall, induce high relative errors in the low flows. Also, in the absence of an infiltration mechanism, all the rainfall is converted into surface runoff by the model. These deficits lead to peaky errors in low flows every once in a while. As these errors will have very low probability to occur, in case of an AR1 process, they get penalized heavily, distorting the parameter inference. Hence, it is recommended, before choosing the marginal of this coupla-based probability model, to ascertain the regions of the hydrograph where high errors can be tolerated and where relatively better predictive capacity of the model is desired. This way during inference, the likelihood function will not produce low probability values for parameters that perform poorly only in low flows but perform adequately well for other sections of the hydrograph.

In case of systematic model deficits, for example Fig. 5, the AR1 likelihood function tends to either assume that the observations have been produced by a process with high standard deviation of errors and high autocorrelation, instead of a process that has relatively low autocorrelation. To mitigate the effect of such interaction between the parameters, we assume that low and the high flows are generated by different marginal, and demonstrate that such an assumption improves parameter estimation. Such flexibility allows the error model to ascribe a different variance to the base flows, as opposed to the high flows, when there is rainfall (Fig. 6).

We also find that non-negativity in the predictions easily achieved using the formulation of marginals used in this work. Even though there are substantial deficits in the low flows, we know that the observations cannot be negative. Therefore, the errors are generally from a skewed distribution. The use of shape parameter as 1 in our case reflects that the error tend to be skewed in low flows, and the model, as in our case, is more likely to overestimate the flow than to underestimate it.

Regarding the robustness of this error model formulation, even though this description is flexible, and can be used to formulate more sophisticated error models, with different marginals, and still be able to capture the autocorrelation, the flexibility is not a guarantee for convergence to meaningful posteriors. The description of the error model still needs to incorporate the relevant knowledge of the modelers about the kind of deficits they expect in their models. Also, the method can perform poorly in cases where the autocorrelation parameter and the heteroscedasticity parameters do not have proper prior constraints. There is no guarantee that a small approximation in the choice of copula or marginal will not diverge the inference from parameter spaces that actually reflect the hydrology of the modeled system. To attain this robustness in inference using copula-based likelihood function, more research is needed.

5 Conclusions

Biases in parameter and predictive uncertainty estimation arise from poorly modelled errors in hydrologic time series. We show that the parameter inference can be skewed in many situations if we always stick to the state-of-the-art AR1 process as the error model. To curtail that we propose a flexible formulation of the likelihood function that allows for different choices of marginal for the time series and takes care of the autocorrelation using copulas. From the results and discussion we conclude:

1. Capturing both non-negativity of the hydrologic system response and the autocorrelation in time is not straightforward.
2. We can use copulas with non-negative marginals to define a probability model for hydrologic time series.
3. This description allows for more flexibility in choosing the marginals. For example, we can have different error structures for low and high flows.
4. If the right marginal densities are chosen, we show that this likelihood formulation provides better parameter and predictive estimates, where additive error models would otherwise perform poorly.
5. However, if the copula likelihood function is not fully reflecting the underlying data generating process, for example, by choosing wrong marginals or wrong heteroscedastic dependence, or having uninformative priors, we again run into issues of non-robustness during inference.
6. Just like in the case of AR1 probability model, the posterior can converge to unrealistic parameter spaces.

This research extends the suite of mathematical descriptions available in hydrology to model time series. Given that such a description is flexible, we foresee that it can be useful for many problems related to parameter inference and model prediction. The analysis is intended to add to the body of literature on representative likelihood functions. It brings forth some benefits and challenges of using a copula-based likelihood function.

Author contributions: OW, with contributions from AS, FC and JR, wrote and structured the main text. OW, with inputs from AS, GE, FC, and JR, conceptualized the methodology. OW, AS, FC, and GE ran simulation experiments. JR and AS supervised the research. All authors reviewed the paper.

Acknowledgement: The authors thank Lorenz Ammann and Peter Reichert for many fruitful discussions. We also thank Miguel Rico-Ramirez for his support during the stay at the University of Bristol, UK. Besides, we are grateful to the Environment Agency and the Centre for Ecology and Hydrology, UK, for the rain gauge and river gauge data respectively. Finally, we thank the editors and the reviewers for their contribution. Part of this study was supported by the EC FP7 project QUICS (Quantifying Uncertainty in Integrated Catchment Studies), grant agreement no. 607000.

References

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

Bárdossy, A., Hörning, S., 2016. Gaussian and non-Gaussian inverse modeling of groundwater flow using copulas and random mixing. Water Resources Research 52 (6), 4504–4526.

Betterle, A., Schirmer, M., Botter, G., 2017. Characterizing the spatial correlation of daily streamflows. Water Resources Research 53 (2), 1646–1663.

Borgomeo, E., Pflug, G., Hall, J. W., Hochrainer-Stigler, S., 2015. Assessing water resource system vulnerability to unprecedented hydrological drought using copulas to characterize drought duration and deficit. Water Resources Research 51 (11), 8927–8948.

Chiles, J.-P., Delfiner, P., 2009. Geostatistics: modeling spatial uncertainty. Vol. 497. John Wiley & Sons.

Del Giudice, D., Albert, C., Rieckermann, J., Reichert, P., 2016. Describing the catchment-averaged precipitation as a stochastic process improves parameter and input estimation. Water Resources Research.

Del Giudice, D., Honti, M., Scheidegger, A., Albert, C., Reichert, P., Rieckermann, J., 2013. Improving uncertainty estimation in urban hydrological modeling by statistically describing bias. Hydrology and Earth System Sciences 17 (10), 4209–4225.

Dogulu, N., López López, P., Solomatine, D. P., Weerts, A. H., Shrestha, D. L., 2015. Estimation of predictive hydrologic uncertainty using the quantile regression and UNEEC methods and their comparison on contrasting catchments. Hydrology and Earth System Sciences 19 (7), 3181–3201.

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 Resources Research 49 (7), 4518–4524.

Gupta, H. V., Sorooshian, S., Yapo, P. Y., 1998. Toward improved calibration of hydrologic models: Multiple and noncommensurable measure of information. Water Resources Research 34 (4), 751–763.

Hall, J. W., Manning, L. J., Hankin, R. K., 2011. Bayesian calibration of a flood inundation model using spatial data. Water Resources Research 47 (5), 1–14.

Honti, M., Stamm, C., Reichert, P., 2013. Integrated uncertainty assessment of discharge predictions with a statistical error model. Water Resources Research 49 (8), 4866–4884.

Kavetski, D., 2018. Parameter Estimation and Predictive Uncertainty Quantification in Hydrological Modelling. In: Duan, Q., Pappenberger, F., Thielen, J., Wood, A., Cloke, H. L., Schaake, J. C. (Eds.), Handbook of Hydrometeorological Ensemble Forecasting. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1–42.

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

Klein, B., Meissner, D., Kobialka, H. U., Reggiani, P., 2016. Predictive uncertainty estimation of hydrological multi-model ensembles using pair-copula construction. Water (Switzerland) 8 (4).

Krzysztofowicz, R., 2001. The case for probabilistic forecasting in hydrology. Journal of Hydrology 249 (1-4), 2–9.

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. Journal of Hydrology 331 (1-2), 161–177.

Liu, Z., Guo, S., Xiong, L., Xu, C. Y., 2018. Hydrological uncertainty processor based on a copula function. Hydrological Sciences Journal 63 (1), 74–86.
URL `https://doi.org/10.1080/02626667.2017.1410278`

Loaiza-Maya, R., Smith, M. S., Maneesoonthorn, W., 2018. Time series copulas for heteroskedastic data. Journal of Applied Econometrics 33 (3), 332–354.
URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.2610

McInerney, D., Thyer, M., Kavetski, D., Lerat, J., Kuczera, G., mar 2017. Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors. Water Resources Research 53 (3), 2199–2239.
URL http://doi.wiley.com/10.1002/2016WR019168

Pianosi, F., Raso, L., 2012. Dynamic modeling of predictive uncertainty by regression on absolute errors. Water Resources Research 48 (3), 1–11.

Refsgaard, J. C., van der Sluijs, J. P., Højberg, A. L., Vanrolleghem, P. A., 2007. Uncertainty in the environmental modelling process - A framework and guidance. Environmental Modelling and Software 22 (11), 1543–1556.

Sadegh, M., Ragno, E., Aghakouchak, A., 2017. Water Resources Research, 5166–5183.

Salvadori, G., De Michele, C., 2004. Frequency analysis via copulas: Theoretical aspects and applications to hydrological events. Water Resources Research 40 (12), 1–17.

Scheidegger, A., 2018. adaptMCMC: Implementation of a Generic Adaptive Monte Carlo Markov Chain Sampler. R package version 1.3.
URL https://CRAN.R-project.org/package=adaptMCMC

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 Resources Research 46 (10), 1–17.

Sun, R., Yuan, H., Liu, X., 2017. Effect of heteroscedasticity treatment in residual error models on model calibration and prediction uncertainty estimation. Journal of Hydrology 554, 680–692.
URL https://doi.org/10.1016/j.jhydrol.2017.09.041

Verkade, J. S., Werner, M. G., 2011. Estimating the benefits of single value and probability forecasting for flood warning. Hydrology and Earth System Sciences 15 (12), 3751–3765.

Vihola, M., 2012. Robust adaptive Metropolis algorithm with coerced acceptance rate.

Wani, O., Beckers, J. V., Weerts, A. H., Solomatine, D. P., 2017. Residual uncertainty estimation using instance-based learning with applications to hydrologic forecasting. Hydrology and Earth System Sciences 21 (8), 4021–4036.

Weerts, A. H., Winsemius, H. C., Verkade, J. S., 2011. Estimation of predictive hydrological uncertainty using quantile regression: Examples from the National Flood Forecasting System (England and Wales). Hydrology and Earth System Sciences.