There are no hydrological monsters, just models and observations with large uncertainties!

George Kuczera[1], **Benjamin Renard**[2], **Mark Thyer**[1] **& Dmitri Kavetski**[1]

[1] *School of Engineering, University of Newcastle, New South Wales 2308, Australia*
george.kuczera@newcastle.edu.au

[2] *Cemagref, UR HHLY, Hydrology-Hydraulics, 3 bis quai Chauveau – CP 220, F-69336 Lyon, France*

Abstract Catchments that do not behave the way the hydrologist expects, expose the frailties of hydrological science, particularly its unduly simplistic treatment of input and model uncertainty. A conceptual rainfall–runoff model represents a highly simplified hypothesis of the transformation of rainfall into runoff. Sub-grid variability and mis-specification of processes introduce an irreducible model error, about which little is currently known. In addition, hydrological observation systems are far from perfect, with the principal catchment forcing (rainfall) often subject to large sampling errors. When ignored or treated simplistically, these errors develop into monsters that destroy our ability to model certain catchments. In this paper, these monsters are tackled using Bayesian Total Error Analysis, a framework that accounts for user-specified sources of error and yields quantitative insights into how prior knowledge of these uncertainties affects our ability to infer models and use them for predictive purposes. A case study involving a catchment with an apparent water balance anomaly (a hydrological monstrosity!) illustrates these concepts. It is found that, in the absence of additional information, the rainfall–runoff record is insufficient to explain this anomaly – it could be due to a large export of groundwater, systematic overestimation of catchment rainfall of the order of 40%, or a conspiracy of these factors. There is "no free lunch" in hydrology. The rainfall–runoff record on its own is insufficient to decompose the different sources of uncertainty affecting calibration, testing and prediction, and hydrological monstrosities will persist until additional independent knowledge of uncertainties is obtained.

Key words Bayesian total error analysis; model structural error; data errors; rainfall–runoff models; ill-posedness

Il n'ya pas de monstres hydrologiques, juste des modèles et des observations avec de grandes incertitudes!

Résumé Les bassins versants se comportant de manière inattendue aux yeux de l'hydrologue mettent en avant les faiblesses des sciences hydrologiques, en particulier le traitement indûment simplifié des incertitudes affectant les données d'entrée du modèle et le modèle lui-même. Un modèle conceptuel est une hypothèse extrêmement simplifiée sur la nature de la transformation pluie–débit. La variabilité spatiale des phénomènes associée à d'inévitables erreurs de conceptualisation des processus conduisent à une erreur structurelle sur le modèle, qui reste encore aujourd'hui largement incomprise. De plus, les appareils et réseaux de mesures hydrologiques sont loin d'être parfaits. En particulier, la principale donnée d'entrée du modèle (la pluie de bassin) est souvent affectée par des erreurs significatives liées à la faible densité spatiale du réseau pluviométrique. Lorsque ces erreurs sont ignorées ou traitées de manière inadéquate, elles se métamorphosent en Monstres maléfiques dont le but est de détruire notre capacité à modéliser le bassin versant. Dans cet article, nous nous attaquons à ces Monstres armés de la méthode BATEA (Bayesian Total Error Analysis). BATEA prend en compte les sources d'erreurs spécifiées par l'hydrologue, et permet d'évaluer de manière quantitative l'impact de notre connaissance a priori de ces incertitudes sur le calage du modèle et son pouvoir prophétique (prédictif). Un bassin versant présentant un bilan en eau aberrant est utilisé pour illustrer ces concepts. Ce cas d'étude montre qu'en l'absence d'informations complémentaires, les données de pluie et de débit ne suffisent pas à lever le voile sur les raisons de cette anomalie. Cette dernière pourrait s'expliquer par un transfert d'eau souterrain important vers des bassins mitoyens, par une sur-estimation systématique d'environ 40% de la pluie de bassin, ou par une conspiration de ces facteurs. En hydrologie, "on n'a rien sans rien": les données de pluie et de débit ne sont pas suffisantes pour démêler les diverses sources d'incertitudes qui affectent le calage, l'évaluation et les prédictions du modèle. Seul une

2

connaissance *a priori* sur ces incertitudes permettra de vaincre les monstres hydrologiques.

Mots clefs analyse Bayesienne de l'erreur totale; erreur structurelle du modèle; erreurs d'observation; modèles pluie–débit; inférence mal posée

1 INTRODUCTION

The proposition that there are hydrological monsters raises several challenges for hydrologists. Monster catchments can be thought of as "deviants" in the sense that their behaviour is far from what the hydrologist expects. As social beings, we label monsters so we can exclude or marginalise them. However, as scientists we should embrace monster catchments. They do us a service in exposing the frailties of our science and the limitations of our world view.

A major issue confronting hydrology is making sense of why, for certain catchments, our model predictions are so poor that in despair we label the catchments (but seldom the models or the methods used to estimate them) as "monsters' (Andreassian *et al*., 2010). It is the goal of this paper to explore how we can make sense of poor model performance. In particular, we argue that hydrologists need a framework that: (i) integrates the process model (here, the conceptual rainfall–runoff – CRR – model) and the observation system to understand the reasons for poor performance, and (ii) can be used to diagnose deficiencies and identify ways for improving the CRR model and the observation system. We also demonstrate, using a real-data case study, how a Bayesian framework can be used to investigate some of these issues.

2 MODEL WRONGNESS AND HYDROLOGICAL MONSTROSITY

A CRR model represents a hypothesis of how rainfall is transformed into runoff:

$$q_t = H(R_t, \eta) \tag{1}$$

The vector q_t denotes the responses of the catchment at time step t. In general, responses are observable point or spatially/temporally averaged quantities, e.g. in the simplest case, the streamflow at the catchment outlet. $H()$ is the hypothesized CRR model, which maps the catchment inputs into its outputs. The history of inputs up to t is denoted by $R_t = \{r_k, k = 1, \ldots, t\}$, where r_k is the input at time step k. Here, r_k contains one or more observable point or spatially/temporally-averaged quantities at the kth time step, typically rainfall and potential evapotranspiration. Note the notational distinction between the upper-case R_t for the history over t time steps, *versus* the lower-case r_k for the quantity at the kth time step. Finally, the vector η contains the model parameters, which complete the model specification and yield the response q_t for a given history of inputs R_t.

Many CRR models have been proposed in the hydrological literature (for an overview, see Singh & Frevert, 2002). Box & Draper (1987, p 424) observed that "all models are wrong, but some are useful". No doubt CRR models are not exempt. But what exactly do we mean when we say a model is "wrong"?

While Box & Draper's definition appears to use "wrong" in the sense of "does not fully represent all complexity of the true system", an arguably more practical view is to label a model as "wrong" if it cannot be reconciled with catchment observations to a degree deemed sufficient for scientific or engineering purposes. However, does encountering a monster catchment necessarily mean the CRR model is wrong in this sense? The real monster may actually be the catchment modelling framework and the observational data (or lack thereof) that are used to calibrate and test models, while the CRR model (and modeller) is merely its victim.

3 UNDERSTANDING MODEL WRONGNESS

When making an assessment of the "wrongness" of a CRR model, a systematic approach is needed to identify all the factors potentially affecting the predictive ability of the model under scrutiny. Crucially, in addition to the CRR model, one needs to characterize the observation system that is responsible for providing the data used to judge the predictive performance of the model. In the absence of an understanding of errors in the observed data, a CRR model may wrongfully be

declared "wrong". Moreover, whether or not one accepts that "all models are wrong", we need a way of characterizing model wrongness or, more precisely, model error.

The error processes affecting each component of the model and observation system are now reviewed.

3.1 Input observation system errors

The catchment responds to external forcing, primarily rainfall and potential evapotranspiration. The true input r_t is estimated using an observation system. In the case of rainfall, the observation system may include one or more raingauges and remote sensing systems such as weather radar. This observation system produces observations $\tilde{r}_t$ that are processed using additional models to make an inference about the true input r_t. Since the observation system is imperfect, there will be uncertainty in the estimated true input r_t. From a Bayesian perspective, this uncertainty can be described using a probability distribution function (pdf) $p(r_t \mid \tilde{r}_t, \theta_r)$, which describes what is known about r_t given the observations $\tilde{r}_t$, the models used to process the observations, and the hyperparameters θ_r defining these models.

Accurate estimation of inputs r_t can be extremely challenging, especially for catchment-average (in time and space) rainfall. For example, Linsley & Kohler (1988) show, using a 1939 study of the Muskingum basin, that large errors in catchment rainfall are the norm rather than the exception. For a 1000-km^2 catchment, a single gauge yielded catchment rainfall estimates with a standard error exceeding 25%. Even if three gauges were used in the averaging, the standard error was only reduced to 16%. Moreover, the smaller the catchment, the larger is the standard error for a given gauge density. These are very sobering statistics. Intuitively, and especially if r_t is estimated with such poor accuracy and precision, knowledge of this uncertainty must be incorporated into the assessment of model wrongness. It is one of the major shortcomings of catchment hydrology that errors in the inputs are typically ignored or treated unduly simplistically. It is likely that, in many cases, this is the birthplace of our "hydrological monsters" (see also Kavetski *et al.*, 2002).

3.2 Output observation system errors

The response of the catchment q_t is monitored using an observation system (e.g. stage measuring equipment), which collects observations ($\tilde{q}_t$) that are interpreted (e.g. using rating curves) to estimate the response q_t. Again, because the observation system is imperfect, there will be uncertainty in the true output q_t, which is summarized by the pdf $p(q_t \mid \tilde{q}_t, \theta_q)$, where θ_q are the hyperparameters defining the model used to process the observations $\tilde{q}_t$.

3.3 Model structural error

Even if there were no errors in the observed forcings and responses, it is likely that CRR models will always contain some irreducible model structural error (e.g. Beven & Binley, 1992; Kuczera *et al.*, 2006). CRR models typically route water through one or more conceptual storages. These one-dimensional (1-D) stores represent 2-D or 3-D features of the catchment and therefore the contents of these conceptual stores are necessarily spatially averaged. The flux of water entering and leaving a store is determined either by forcing inputs, or by flux equations conditioned on the contents of the store and parameter values. If any of these fluxes is in error, it is likely that the catchment response will be in error.

However, even if the observed forcings are error-free, the fact that they are spatial and temporal averages of random fields introduces an irreducible model error, often referred to as "sub-grid variability" (e.g. Andreassian *et al.*, 2004; Ching *et al.*, 2006). There are infinitely many spatially and temporally distributed rainfall fields that yield the same average catchment rainfall. However, each distinct rainfall field causes a different hydrological response. For example, if the main mass of the rainfall field were located over the saturated part of the catchment, a significant quickflow response would occur. Conversely, if that same rainfall were located over an

unsaturated part of the catchment, no quickflow response would occur and the soil store would be recharged.

In a similar vein, if inputs are temporally averaged over scales exceeding catchment response times, unavoidable errors are introduced. For example, consider the widespread practice of using daily rainfall inputs. The same daily rainfall depth could be uniformly distributed over the day, or concentrated in an intense burst. These scenarios produce very different quickflow responses, often on time scales considerably less than a day. It follows that models that utilize spatially and temporally averaged input fields and storages are fundamentally incapable of accurately simulating catchment response – such models are, in this sense, intrinsically "wrong".

These considerations suggest that a CRR model cannot be deterministic, in the sense that true catchment response could not be reproduced even if the true parameters and exact spatially and temporally averaged forcing were known. Characterizing intrinsic model structural error probabilistically, one can begin formulating hypotheses about its properties. While there is currently no wide agreement on how best to describe model structural error, several options are available. What is proposed here is a working hypothesis based on concepts developed in state-space modelling (e.g. Bras & Rodriguez-Iturbe, 1984)

A probabilistic model error can be introduced into a deterministic CRR model in at least two ways:

(1) Exogenous random error. Huard & Mailhot (2008) use an exogenous random term to describe the structural error of the CRR model. The simplest case is the additive Gaussian error:

$$q_t \leftarrow N\left(q \,|\, H(R_t, \eta), \theta_m^2\right) \qquad (2)$$

where $N(x/m, s^2)$ denotes that x is a Gaussian random variable with mean m and variance s^2. This approach assumes that, on average, the deterministic CRR model $H(R_t, \eta)$ correctly predicts the true response q_t and that its errors are Gaussian with constant variance θ_m^2.

The additive error model (2) is probably too simple to fully account for model structural error. Therefore we use it to represent "remnant error" that attempts to capture, at least approximately, errors ignored elsewhere (see Renard *et al*., 2010, for a detailed discussion). The magnitude of θ_m^2 can be viewed as a measure of our success in accounting for all sources of uncertainty – the smaller θ_m^2, the more successful is the attempt.

(2) Introduce one or more stochastic CRR parameters. Since many of the factors causing model structural error are identifiable, one can be more specific in the choice of mechanisms for introducing randomness into the CRR model than additive Gaussian noise. For example, Kuczera *et al.* (2006) argued that the mass balance in CRR models should be maintained, and that random errors due to processes such as sub-grid variability should manifest themselves as random perturbations in one or more fluxes between CRR storage elements. This can be accomplished by partitioning the CRR model parameters η into two groups: deterministic parameters ω and stochastic (time-varying) parameters λ.

The stochastic parameters ("latent variables" in Bayesian terminology) are sampled from a stochastic process:

$$\lambda_t \leftarrow p(\lambda \,|\, \Lambda_{t-1}, \theta_\lambda) \qquad (3)$$

where Λ_{t-1} is the history of latent variables $\{\lambda_1, ..., \lambda_{t-1}\}$ and θ_λ is the vector of hyper-parameters defining the stochastic model of λ. Currently, little is known about formulating $p(\lambda \,|\, \Lambda_{t-1}, \theta_\lambda)$ to reflect sub-grid variability; this remains a topic for future work.

4   BAYESIAN INFERENCE FRAMEWORK

Having discussed the probabilistic representation (hypothesis) of model and observation system errors, we now consider a framework that uses these hypotheses to gain insights into the performance of a CRR model. The conclusions are necessarily conditioned on available data,

namely the time series of observed inputs $\tilde{R} = \{\tilde{r}_t, t = 1,..,N_t\}$ and outputs $\tilde{Q} = \{\tilde{q}_t, t = 1,..,N_t\}$ of length N_t, as well as on other prior information and on assumptions regarding the CRR model and data uncertainty and structural error models.

We use the Bayesian Total Error Analysis (BATEA) framework (see Kavetski *et al.*, 2002, 2006a,b; Kuczera *et al.*, 2006, for derivation; Thyer *et al.*, 2009, for analysis of parameter consistency and posterior diagnostics; and Renard *et al.*, 2010, for identifiability evaluation). The basic BATEA principle is to use all available information when inferring quantities of interest. Such quantities might include: the deterministic CRR parameters ω, the hyperparameters θ_λ defining the stochastic model $p(\lambda \mid \Lambda_{t-1}, \theta_\lambda)$, the actual values of the latent variables $\Lambda = \{\lambda_1,...,\lambda_n\}$, the remnant model error variance θ_m^2, the true rainfall history $R = \{r_t, t = 1,..,N_t\}$, the hyperparameters θ_r defining $p(R \mid \tilde{R}, \theta_r)$, and the hyperparameters θ_q defining response errors $p(q_t \mid \tilde{q}_t, \theta_q)$. In this study, the response error hyperparameters θ_q are assumed to be known (e.g. they can be determined from rating curve analysis, see Thyer *et al.*, 2009). This assumption can be relaxed if necessary.

Using Bayes' theorem, the posterior pdf of $\{R, \theta_r, \omega, \theta_\lambda, \Lambda, \theta_m\}$ is derived in equation (4)

$$
\begin{aligned}
&p(R, \theta_r, \omega, \theta_\lambda, \Lambda, \theta_m \mid \tilde{Q}, \tilde{R}, \theta_q) \\
&= \frac{p(\tilde{Q} \mid R, \theta_r, \omega, \theta_\lambda, \Lambda, \tilde{R}, \theta_m, \theta_q)\, p(R, \theta_r, \omega, \theta_\lambda, \Lambda, \tilde{R}, \theta_m \mid \theta_q)}{p(\tilde{Q}, \tilde{R} \mid \theta_q)} \\
&\propto p(\tilde{Q} \mid R, \omega, \Lambda, \theta_m, \theta_q)\, p(\omega)\, p(\theta_m)\, p(\Lambda, \theta_\lambda)\, p(R, \tilde{R}, \theta_r) \\
&\propto p(\tilde{Q} \mid R, \omega, \Lambda, \theta_m, \theta_q)\, p(\omega)\, p(\theta_m)\, p(\Lambda \mid \theta_\lambda)\, p(\theta_\lambda)\, p(R \mid \tilde{R}, \theta_r)\, p(\theta_r \mid \tilde{R})\, p(\tilde{R}) \\
&\propto \underbrace{p(\tilde{Q} \mid R, \omega, \Lambda, \theta_m, \theta_q)}_{\text{likelihood on response}} \times \underbrace{p(\omega)}_{\substack{\text{prior on deterministic} \\ \text{CRR parameters}}} \times \underbrace{p(\theta_m)}_{\substack{\text{prior on remnant model} \\ \text{error variance}}} \\
&\times \underbrace{p(\Lambda \mid \theta_\lambda)}_{\substack{\text{pdf of stochastic} \\ \text{CRR latent variables}}} \times \underbrace{p(\theta_\lambda)}_{\substack{\text{prior on stochastic CRR} \\ \text{hyperparameters}}} \times \underbrace{p(R \mid \tilde{R}, \theta_r)}_{\substack{\text{pdf of true inputs} \\ \text{given observed inputs}}} \times \underbrace{p(\theta_r)}_{\substack{\text{prior on input error} \\ \text{hyperparameters}}}
\end{aligned}
\tag{4}
$$

where the simplifications in the second and subsequent lines arise from conditional independence between random variables and from combining the pdfs involving the data $\{\tilde{R}, \tilde{Q}\}$ into the constant of proportionality.

Formulation (4) closely follows the original derivation of BATEA (Kavetski *et al.*, 2002) and directly includes the true input as the subject of inference. It is equivalent to the derivations by Kavetski *et al.* (2006a) and Kuczera *et al.* (2006), but has a clearer interface for using prior information on the inputs (Renard *et al.*, 2009).

Several terms on the right hand side of (4) deserve further comment:

(1) The likelihood function $p(\tilde{Q} \mid R, \omega, \Lambda, \theta_m, \theta_q)$ is the sampling distribution of the observed response for a given set of latent variables and true inputs. It incorporates the effects of errors in the observed response and remnant model errors.

The likelihood can be obtained using total probability integration over the unknown true response Q (necessarily using an estimate of its distribution, e.g. from rating curve analysis, Thyer *et al.*, 2009). For the case of independent homoscedastic remnant errors $N\left(q_t \mid m = H(R_t, \omega, \lambda_t), s^2 = \theta_m^2\right)$ and independent heteroscedastic response errors $N\left(\tilde{q}_t \mid m = q_t, s^2 = \left(H(R_t, \omega, \lambda_t) \times \theta_q\right)^2\right)$, the likelihood is given by the multi-dimensional

integral:

$$p(\tilde{Q} \mid R, \omega, \Lambda, \theta_m, \theta_q) = \iint p\left(\tilde{Q} \mid Q, R, \omega, \Lambda, \theta_q\right) p\left(Q \mid R, \omega, \Lambda, \theta_m\right) \mathrm{d}Q$$

$$= \prod_{t=1}^{n} \int N\left(\tilde{q}_t \mid m = q_t, s^2 = \left(H(R_t, \omega, \Lambda_t) \times \theta_q\right)^2\right) N\left(q_t \mid m = H(R_t, \omega, \Lambda_t), s^2 = \theta_m^2\right) \mathrm{d}q_t \qquad (5)$$

$$= \prod_{t=1}^{n} N\left(\tilde{q}_t \mid m = H(R_t, \omega, \Lambda_t), s^2 = \theta_m^2 + \left(H(R_t, \omega, \Lambda_t) \times \theta_q\right)^2\right)$$

Equation (5) states that observed responses are described by a Gaussian distribution centred on the CRR model prediction $H(R_t, \omega, \lambda_t)$, with a variance equal to the sum of the (homoscedastic) remnant error variance θ_m^2 and the (heteroscedastic) observation error variance $\left(H(R_t, \omega, \lambda_t) \times \theta_q\right)^2$.

(2) The pdf $p(\omega)$ describes the prior information on the deterministic CRR parameters. It may be non-informative, but ideally should take advantage of regional transfer of information using catchment attributes, previous calibrations of the CRR model, etc.

(3) In contrast, little is known a priori about the structural-error hyperparameters θ_λ. Consequently, the pdf $p(\theta_\lambda)$ is likely to be non-informative in most applications.

(4) The pdf $p(R \mid \tilde{R}, \theta_r)$ describes what is known about the true inputs given input observations. The specification of this pdf is critical to the well-posedness of the posterior (4) (Renard *et al.*, 2010).

(5) The pdf $p(\theta_r)$ describes the prior information on the parameters defining the input observation model. Ideally this would be derived from analysis of rainfall data sets sufficiently rich to infer the parameters in a meaningful way.

The BATEA posterior (4) reduces to the traditional least squares and Nash-Sutcliffe criteria commonly used in CRR model calibration under the following conditions: (a) there are no rainfall errors; (b) all CRR parameters are deterministic; (c) the runoff error has constant variance; and (d) priors on all quantities of inference are non-informative.

The posterior (4) appears disarmingly simple. Yet in truth, it is an inferential "beast" feasting on the observed data, devouring the modeller's assumptions about the rainfall–runoff process and errors, and generously "returning" a digested summary of what can be learned about $\{R, \theta_r, \omega, \theta_\lambda, \Lambda, \theta_m\}$. It is a powerful tool to analyse the impact of data and model uncertainty on the model parameters and predictions.

Ideally we would like to achieve a well-posed inference of $\{R, \theta_r, \omega, \theta_\lambda, \Lambda, \theta_m\}$ and meaningful scrutiny of the assumptions underpinning the posterior (Gelman *et al.*, 2004; Thyer *et al.*, 2009). However, as we shall see, there is "no free lunch". Well-posed inference is only possible if there is sufficient independent information.

5   CASE STUDY: A TALE OF A MONSTER CATCHMENT

The case study explores a "monster" catchment, whose "monstrosity" is rooted in a water balance anomaly (Le Moine *et al.*, 2007). The case study illustrates the hydrological insights yielded by a careful application of BATEA and shows that the catchment is not necessarily "monstrous" *per se*, but rather, that there is insufficient information to resolve its water balance anomaly.

5.1 The catchment and the model

The case study is based on the 156 km^2 Vesonne catchment located in the Rhone-Alps region of France. Its average daily statistics are: rainfall of 2.87 mm/d; potential evapotranspiration of 2.29 mm/d and runoff of 0.13 mm/d. The selected CRR model is GR4J (Perrin *et al.*, 2003), which has four parameters: production store capacity X1, water exchange coefficient X2, routing store

capacity X3 and routing time parameter X4. A particular feature of GR4J is its groundwater import/export, which is controlled by parameter X2.

5.2 Least squares calibration

The GR4J model was calibrated to a two-year daily rainfall–runoff series (from 8 September 1999 to 8 August 2001) using the least squares criterion, yielding a Nash-Sutcliffe statistic of 0.65. Figure 1 compares the observed and fitted runoff time series. It can be seen that the SLS-based model predictions contain gross errors for many of the days with large runoff.

Moreover, the posterior distribution of the model parameters, shown in Fig. 2, reveals further troublesome behaviour. When negative, parameter X2 can be interpreted as the maximum daily groundwater export to another catchment or regional aquifer. Figure 2 shows that X2 has a modal (most-likely) value of approx. –33 mm/d. Over the calibration period, this translates into an inferred daily average export of about 2.1 mm/d (which is over 70% of average daily rainfall), indicating that "monstrous" volumes of water are being exported from the catchment, presumably through some unknown groundwater path.

This inference is conditioned on two assumptions embedded in least squares calibration: (i) the rainfall is observed without error, and (ii) model and runoff observation errors can be represented by an independent constant variance process. In reality, both assumptions are flawed. Therefore, the conclusion that the Vesonne has a massive leak rests on assumptions that cannot be defended. The conclusion is itself hence in doubt.

5.3 BATEA calibration with rainfall input multipliers

We relax two assumptions made in the least squares calibration:

(a) It is unrealistic to assume that catchment rainfall is estimated without error. Unfortunately, other than the daily raingauge record, there is no supplementary information on the accuracy and bias of the Vesonne catchment rainfall estimates. As a result, the only way to get information about rainfall errors is to infer them indirectly using the information in the runoff record in conjunction with the GR4J model. We adopt a simple rainfall error model (Kavetski *et al.*, 2006b), which assumes multiplicative lognormal rainfall errors ϕ corrupting the estimated depth of each storm event

$$p(r_t \mid \tilde{r}_t, \theta_r) := \left\{ r_t = \phi_t \tilde{r}_t; \quad \log_e(\phi_t) \leftarrow N(\phi \mid \mu_\phi, \sigma_\phi^2) \right\} \tag{6}$$

In the absence of independent knowledge of rainfall uncertainty, the hyperparameters $\theta_r = \{\mu_\phi, \sigma_\phi\}$ are inferred from the data using a weakly informative prior.

(b) A 10% standard error in the observed runoff is assumed, representing our *a priori* expectation of errors in a typical flow gauging observation system.

As before, the GR4J model is treated as deterministic (i.e. no time-varying parameters were used in its calibration). Moreover, it is also assumed that errors in estimating potential evapotranspiration (PE) are negligible. Indeed, PE exhibits far less variability than rainfall and therefore is considered far less susceptible to the large sampling errors that affect catchment rainfall at daily time scales (e.g. Oudin *et al.*, 2006).

Given these assumptions, BATEA yields the fit in Fig. 3, with a Nash-Sutcliffe statistic of 0.83. The parameter inference summarized in Fig. 4 shows the posteriors of the GR4J parameters have shifted, with the biggest change for parameter X2. Indeed the X2 parameter is virtually zero, suggesting there is no large-scale leakage of water from the Vesonne!

This radical change in conclusion can be understood by considering Fig. 5, which displays the posterior of the rainfall log-multiplier hyperparameters $\{\mu_\phi, \sigma_\phi\}$. The posterior means of the hyperparameters $\{\mu_\phi, \sigma_\phi\}$ correspond to an expected value of the rainfall multipliers ϕ of about 0.57. This implies that, if we allow for rainfall errors, the water balance anomaly could be a result of the raingauge overestimating the actual catchment rainfall by about 43%.

Interestingly, although the parameter X2 was free to vary over a wide range of negative values, the calibration favoured a no-leakage inference and instead explained the anomaly in the catchment water balance by a systematic overestimation of catchment rainfall by the raingauge.

In this scenario, which admits rainfall error but no model error, BATEA suggests the variability in rainfall errors is very large, with Fig. 5 reporting a standard error of 40%! This error may be considered large, but certainly not unrealistic given the results reported by Linsley & Kohler (1988). However, since no explicit provision has been made for model error, there is a distinct possibility that the inferred rainfall multipliers may be accounting for both rainfall and model structural error. This issue is investigated in the following section (see also Renard *et al.*, 2010, for an in-depth discussion).

5.4 BATEA calibration with stochastic X2 parameter

To gain insights into the structural error of the GR4J model, it was assumed that rainfall errors were negligible, and that parameter X2 varies on a storm-event scale and can be described by a lognormal distribution

$$p(\lambda \,|\, \theta_\lambda) := \left\{ \lambda = \text{X2}; \quad \log_e(-\text{X2}) \leftarrow N(\text{X2} \,|\, \mu_{X2}, \sigma^2_{X2}) \right\} \tag{7}$$

The log-normal distribution was chosen to ensure that X2 values are negative, which is consistent with prior analyses, suggesting that groundwater is exported from this catchment (Le Moine *et al.*, 2007). Given this assumption, BATEA yields a fit similar to that shown in Fig. 3. The posterior distribution of the X2 hyperparameters $\{\mu_{X2}, \sigma_{X2}\}$ is displayed in Fig. 6. The posterior of μ_{X2} indicates negative values of the order of -50 mm/day. This conclusion is consistent with that made using the least squares calibration, and is not unexpected – with rainfall assumed error-free, the GR4J model is forced to export water in order to close the catchment water balance.

5.5 BATEA calibration with rainfall multipliers and stochastic X2 parameter

It is almost certain that rainfall data errors coexist with model structural errors. It is therefore logical to evaluate the posterior (4) using both (6) and (7) to make allowance for both rainfall and structural errors, and check whether this affects the interpretation of the catchment water balance anomaly.

However, there is a problem. There are three sources of uncertainty contributing to the total uncertainty in the discharge. In the absence of knowledge about at least two of these sources, the inference on individual sources can become ill-posed, meaning there is insufficient information to make a useful inference. The problem is akin to inferring the water balance of the catchment: if *n* sources are known, then we need independent information on at least *n*–1 before the entire mass balance can be reconciled.

Ill-posed inferences are not always easy to detect and can lead to meaningless predictions, particularly in model extrapolation. Fortunately, the performance of the Markov Chain Monte Carlo algorithm that samples from (4) can be used to detect ill-posedness (Renard *et al.*, 2010). In particular, non-informative posterior distributions and/or failure to converge due to exceedingly strong correlations are sure signs that the inference is ill-posed.

A clear symptom of this ill-posedness is demonstrated in Fig. 9, which shows the absolute posterior correlation when inferring both input and model structural errors without informative priors. The correlation structure is complex, characterized by substantial correlation between concurrent and neighbouring rainfall multipliers ϕ and latent variables X2. The implication is that ϕ and X2 mimic each other (see strong correlations in the off-diagonal bloc of Fig. 7): distinct combinations $\{\phi, \text{X2}\}$ can yield very similar simulated runoff time series. It is important to note that the scenarios in sections 5.3 and 5.4 (which did not attempt to estimate input and model errors concurrently) did not exhibit such correlations and were hence well-posed (but likely inaccurate due to ignoring one of the errors, see Renard *et al.*, 2009, 2010, for background theory on ill-posedness and non-identifiability).

The ill-posedness of the posterior confirms the intuitive notion that, without additional

information, the underlying cause for the water balance anomaly cannot be resolved. Since little is known about the magnitude and properties of model structural error, it is preferable to seek additional information on rainfall uncertainty, e.g. from an upgraded rainfall observation system that reduces sampling error and bias to more accurately and precisely specify $p(r_t \mid \tilde{r}_t, \theta_r)$ (Renard *et al*., 2009, 2010). In addition, more detailed hydrogeologic investigations and regional information on neighbouring catchments could yield insights into whether the groundwater leak is realistic.

Ultimately, the "monstrosity" of the Vesonne cannot be resolved using solely rainfall–runoff data of unspecified quality. Nevertheless, even in such information-poor conditions, BATEA yields useful insights into the magnitude and variability of rainfall errors and model parameters likely to be consistent with the data. At least, this could be viewed as an upper bound on rainfall uncertainty or internal GR4J stochasticity.

These insights are in good agreement with the findings of Le Moine *et al*. (2007), who compared four distinct alternatives to close the water balance of the Vesonne catchment: (i) correcting precipitation, (ii) correcting potential evapotranspiration, (iii) correcting catchment area, and (iv) allowing for groundwater export/import. They found that such corrections produce near-equivalent simulated hydrographs, suggesting that reliance on runoff goodness-of-fit to identify the closure mechanism is fraught with difficulty. The hydrologist has no choice but to rely on other types of information. That said, it is unlikely that a single closure mechanism is dominant. Invariably, input and model structural errors coexist in most practical hydrological contexts. The hydrologist must therefore use all available sources of information to formulate meaningful priors.

CONCLUSIONS

A catchment that cannot be modelled adequately is not necessarily a monster. Indeed, if there are any monsters, we should first look for them in the realm of hydrological science. This paper has argued that the science of catchment hydrology needs a framework that jointly analyses the hydrological model and the observation system that informs it. This framework must then be used to sift through competing hypotheses and the observed evidence to draw trustworthy quantitative inferences and conclusions. This is hardly a new idea – all areas of science carry out such analysis in one form or another – but the problem is that hydrologists, confronted by multiple and substantial sources of uncertainty, are yet to even agree on such a systematic framework, let alone apply it.

The Bayesian Total Error Analysis used in this paper provides a natural means to make inferences and draw conclusions from hydrological data. The BATEA posterior (4) is constructed to account for information about the quality of forcing and response data, and to expose all hypotheses on model dynamics, model structural error and observational data uncertainty to scrutiny.

A core requirement of any inference framework is that it be trustworthy. Trustworthiness should not be confused with accuracy or precision. As illustrated in the Vesonne case study, no meaningful conclusion could be drawn about cause of the water balance anomaly – given that the available data lacked any description of its uncertainties, the problem was ill-posed and it was impossible to discriminate between competing explanations of the water balance anomaly. This is not a failure of the Bayesian inference framework used in this case study, but rather a demonstration of its trustworthiness in quantitatively diagnosing the insufficient informational content of the data and indicating the type of information needed to resolve the ill-posedness.

The conclusions of any statistical inference framework are conditioned on the available information. For the Bayesian framework, these are the data that inform the likelihood function, the prior information on data uncertainties and CRR parameters, as well as the hypothesized model and error structure. It is imperative that all these assumptions be critically scrutinized – in the absence of this scrutiny, the engine of science stalls.

The Vesonne case study illustrates several fundamental properties of the Bayesian posterior. Inferences can be sensitive to the assumptions made about errors – this is the fiefdom of Lord

GIGO (garbage in, garbage out). Indeed, there is "no free lunch". BATEA cannot produce useful inferences if the data are not sufficiently informative. In particular, calibrating a CRR model solely to rainfall–runoff time series with no independent estimates of data and model accuracy yields, at best, some (limited) insight about the total error, but little about its causes and components.

In the future, the use of independently-derived informative priors on rainfall and runoff errors can be used to isolate model structural error and understand its properties. How much prior information is necessary is a topic of our ongoing work. Moreover, it is reasonable to expect that, if model wrongness is to be better understood, catchment observation systems will have to produce rainfall and runoff estimates with sufficiently small sampling errors so that model structural error represents a significant fraction of the total error. Advances in this direction will help us purge hydrological "monsters" and dethrone Lord GIGO who rules them.

REFERENCES

Andréassian, V., Perrin, C., Parent, E. & Bardossy, A. (2010) Court of Miracles of Hydrology: can failure stories contribute to hydrological science? *Hydrol. Sci. J.* (this issue).

Andréassian, V., Oddos, A., Michel, C., Anctil, F., Perrin, C. & Loumagne, C. (2004) Impact of spatial aggregation of inputs and parameters on the efficiency of rainfall–runoff models: A theoretical study using chimera watersheds. *Water Resour. Res.* **40**(5), W05209.

Beven, K. & Binley, A. (1992) The future of distributed models – model calibration and uncertainty prediction. *Hydrol. Processes* **6**(3), 279–298.

Box, G. E. P. & Draper, N. R. (1987) *Empirical Model-Building and Response Surfaces*. Wiley.

Bras, R. L. & Rodriguez-Iturbe, I. (1984) *Random Functions in Hydrology*. Addison-Wesley.

Ching, J., Herwehe, J. & Swall, J. (2006) On joint deterministic grid modeling and sub-grid variability conceptual framework for model evaluation. *Atmos. Environ.* **40**(26), 4935–4945.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2004) *Bayesian Data Analysis*, Second edition. New York: Chapman & Hall/CRC.

Huard, D. & Mailhot, A. (2008) Calibration of hydrological model GR2M using Bayesian uncertainty analysis. *Water Resour. Res.* **44**(2), W02424. doi:10.1029/2007WR005949.

Kavetski, D., Franks, S. W. & Kuczera, G. (2002) Confronting input uncertainty in environmental modeling. In: *Calibration of Watershed Models* (ed. by Q. Duan, H. V. Gupta & S. Sorooshian). AGU Series, vol. 6, 49-68. Washington: American Geophysical Union.

Kavetski, D., Kuczera, G. & Franks, S. W. (2006a) Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. *Water Resour. Res.* **42**(3). W03407, doi:10.1029/2005WR004368.

Kavetski, D., Kuczera, G. & Franks, S. W. (2006b) Bayesian analysis of input uncertainty in hydrological modeling: 2. Application. *Water Resour. Res.* **42**(3). W03408, doi:10.1029/2005WR004376,.

Kuczera, G., Kavetski, D., Franks, S. W. & 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.

Le Moine, N., Andreassian, V., Perrin, C. & Michel, C. (2007) How can rainfall–runoff models handle intercatchment groundwater flows? Theoretical study based on 1040 French catchments. *Water Resour. Res.* **43**, W06428. doi:10.1029/2006WR005608.

Linsley, R. K. & Kohler, M. A. (1988) *Hydrology for Engineers*. London: McGraw Hill.

Oudin, L., Perrin, C., Mathevet, T., Andreassian, V. & Michel, C. (2006) Impact of biased and randomly corrupted inputs on the efficiency and the parameters of watershed models. *J. Hydrol.* **320**(1-2), 62–83.

Perrin, C., Michel, C. & Andreassian, V. (2003) Improvement of a parsimonious model for streamflow simulation, *J. Hydrol.* **279**, 275–289.

Renard, B., Leblois, E., Kuczera, G., Kavetski, D., Thyer, M. & Franks, S. W. (2009) Characterizing errors in areal rainfall estimates: application to uncertainty quantification and decomposition in hydrological modelling. In: *Proceedings of 32nd Hydrology and Water Resources Symposium* (Newcastle, Australia).

Renard, B., Kavetski, D., Kuczera, G., Thyer, M. & Franks, S. W. (2010) Understanding predictive uncertainty in hydrologic modeling: the challenge of identifying input and structural errors. *Water Resour. Res.* (in press).

Singh, V. P. & Frevert, D. K. (2002) *Mathematical Models of Small Watershed Hydrology and Applications*. Water Resources Publications, 972pp.

Thyer, M., Renard, B., Kavetski, D., Kuczera, G., Franks S. W. & Srikanthan, S. (2009) Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: a case study using Bayesian total error analysis. *Water Resour. Res.* **45**, doi:10.1029/2008WR006825.
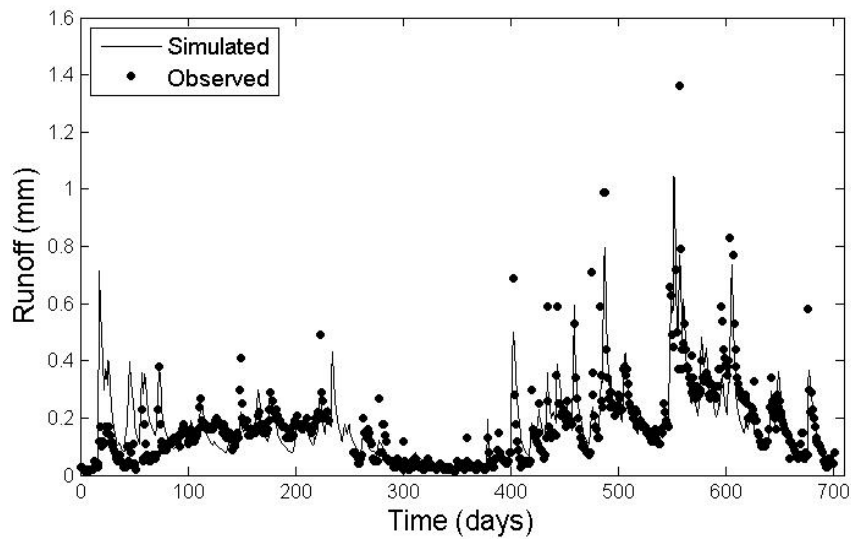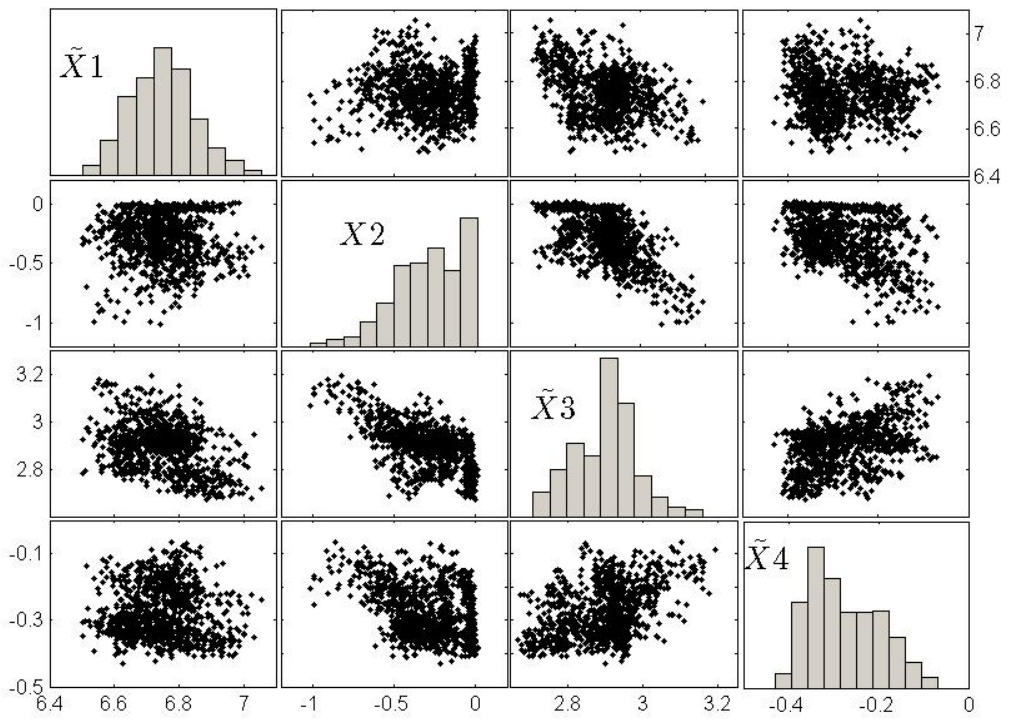
Fig. 1 Time series of observed and daily runoff fitted using least squares (calibration period).
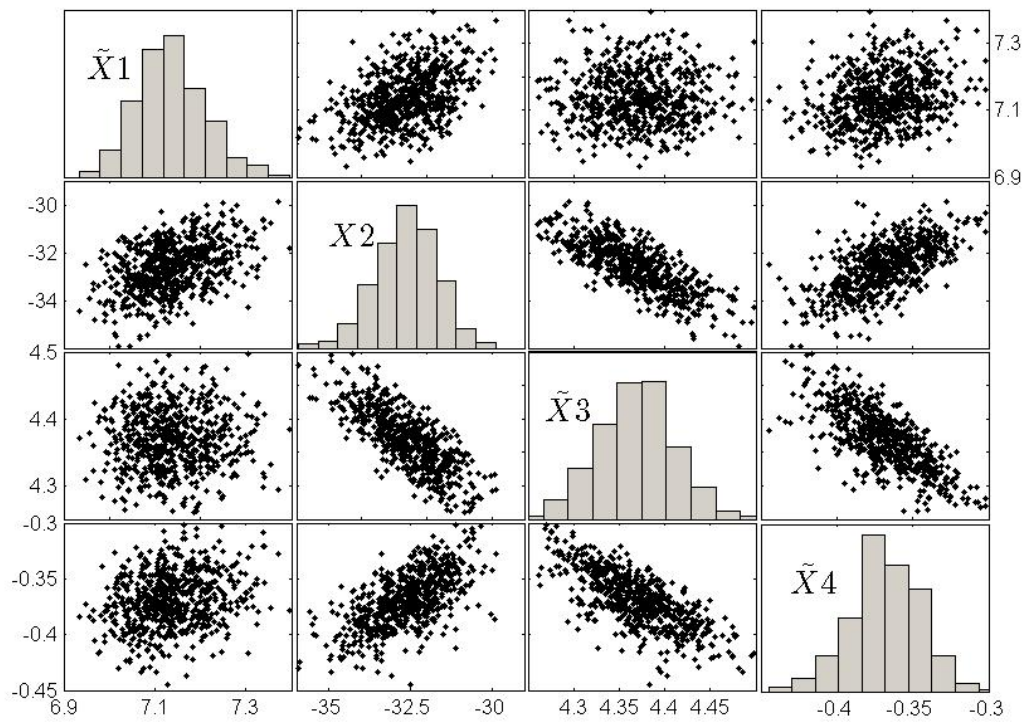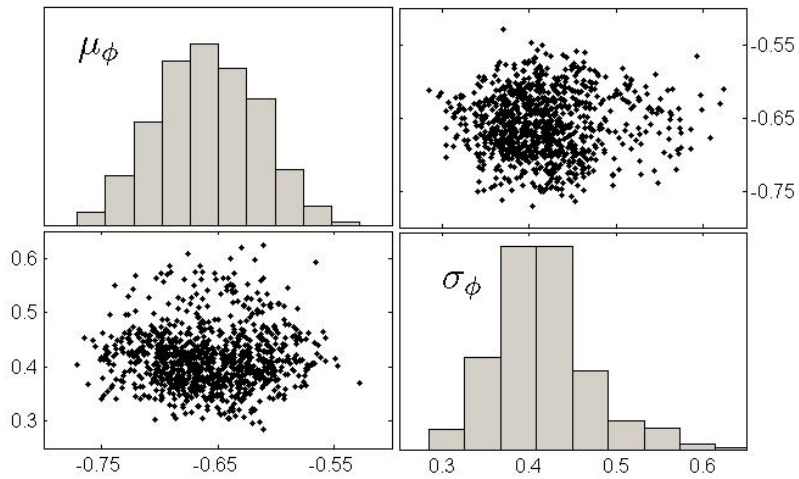

Fig. 2 Posterior distributions of GR4J parameters fitted using least squares. The tilde indicates parameter transformations: $\tilde{X}1 = \log(X1)$, $\tilde{X}3 = \log(X3)$ and $\tilde{X}4 = \log(X4 - \frac{1}{2})$.
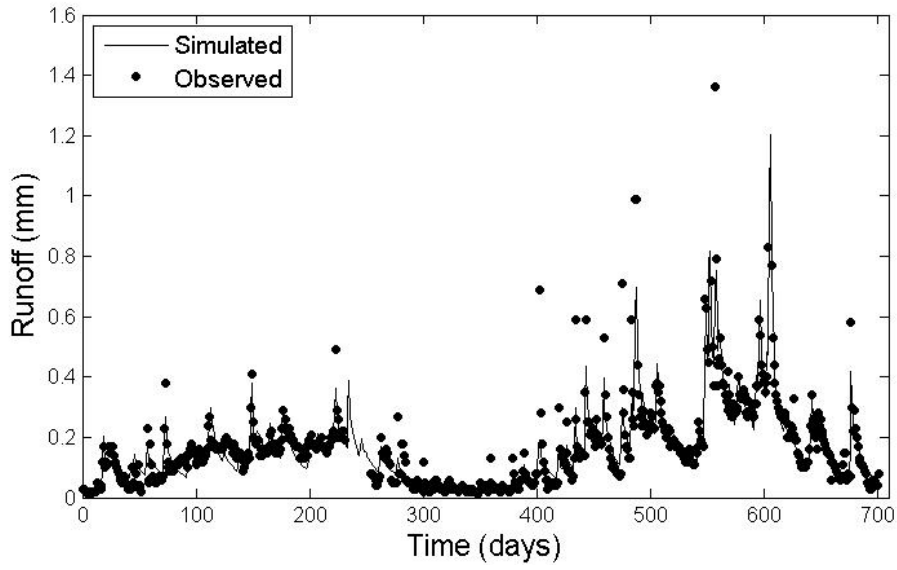
Fig. 3 Observed and predicted runoff fitted using BATEA with rainfall multipliers (calibration period).


Fig. 4 Posterior distributions of GR4J parameters fitted using BATEA with rainfall multipliers. Note the different axis scaling compared to Fig. 3.

Fig. 5 Posterior distributions of rainfall multiplier hyperparameters: top left box shows the mean μ_ϕ and bottom right box shows the standard deviation σ_ϕ
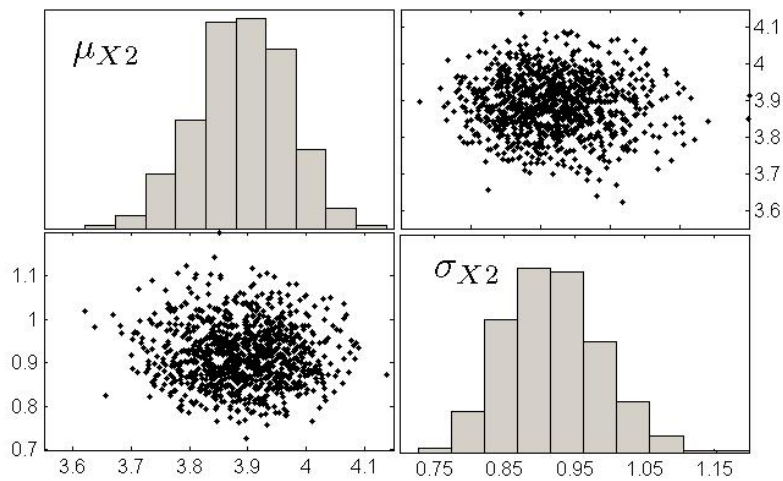
Fig. 6 Posterior distributions of X2 hyperparameters: top left box shows the mean μ_{X2} and bottom right box shows the standard deviation σ_{X2}.
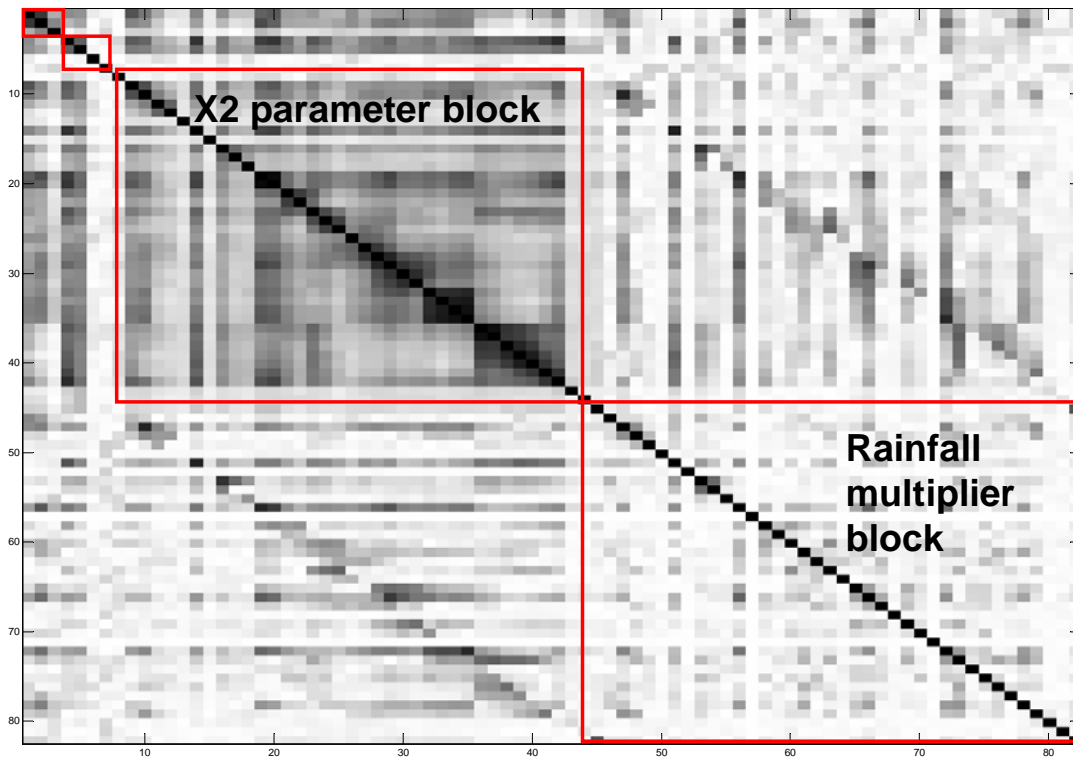
Fig. 7 Absolute posterior correlation between the rainfall multipliers and X2 latent variables (black denotes a perfect correlation of 1, white denotes no correlation).