WATER RESOURCES RESEARCH, VOL. 49, 4335–4345, doi:10.1002/wrcr.20354, 2013

Toward diagnostic model calibration and evaluation: Approximate Bayesian computation

Jasper A. Vrugt[1,2,3] and Mojtaba Sadegh[1]

Received 2 November 2012; revised 25 April 2013; accepted 4 June 2013; published 29 July 2013.

[1] The ever increasing pace of computational power, along with continued advances in measurement technologies and improvements in process understanding has stimulated the development of increasingly complex hydrologic models that simulate soil moisture flow, groundwater recharge, surface runoff, root water uptake, and river discharge at different spatial and temporal scales. Reconciling these high-order system models with perpetually larger volumes of field data is becoming more and more difficult, particularly because classical likelihood-based fitting methods lack the power to detect and pinpoint deficiencies in the model structure. Gupta et al. (2008) has recently proposed steps (amongst others) toward the development of a more robust and powerful method of model evaluation. Their diagnostic approach uses signature behaviors and patterns observed in the input-output data to illuminate to what degree a representation of the real world has been adequately achieved and how the model should be improved for the purpose of learning and scientific discovery. In this paper, we introduce approximate Bayesian computation (ABC) as a vehicle for diagnostic model evaluation. This statistical methodology relaxes the need for an explicit likelihood function in favor of one or multiple different summary statistics rooted in hydrologic theory that together have a clearer and more compelling diagnostic power than some average measure of the size of the error residuals. Two illustrative case studies are used to demonstrate that ABC is relatively easy to implement, and readily employs signature based indices to analyze and pinpoint which part of the model is malfunctioning and in need of further improvement.

Citation: Vrugt, J. A., and M. Sadegh (2013), Toward diagnostic model calibration and evaluation: Approximate Bayesian computation, *Water Resour. Res.*, *49*, 4335–4345, doi:10.1002/wrcr.20354.

1. Introduction and Scope

[2] Consider a discrete vector of measurements, $\tilde{\mathbf{Y}} = \{\tilde{y}_1, \ldots, \tilde{y}_n\}$ observed at times $t = \{1, \ldots, n\}$ that summarizes the response of an environmental system $\Im$ to forcing variables $\mathbf{U} = \{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$. Let $\mathbf{Y} = \{y_1, \ldots, y_n\}$ denote the corresponding predictions from a dynamic (nonlinear) model, f with parameter values θ,

$$\mathbf{Y} \leftarrow f(\mathbf{x}_0, \theta, \tilde{\mathbf{U}}), \tag{1}$$

where $\mathbf{x}_0$ signifies the initial states of the system at $t = 0$. The residual vector defines the difference between the actual and model-simulated system behavior,

$$\mathbf{E}(\theta) = F[\tilde{\mathbf{Y}}] - F[\mathbf{Y}(\theta)] = \{e_1(\theta), \ldots, e_n(\theta)\}, \tag{2}$$

where $F[\cdot]$ allows for transformations of the observations and model predictions. The goal is now to find those values of $\theta \in \Theta \in \mathbb{R}^d$ that best mimic the observed system behavior.

[3] During the past four decades much research has been devoted to the development of computer based methods for fitting hydrologic models to calibration data (e.g., streamflow, water chemistry, groundwater table depth, soil moisture, snow water equivalent). That research has primarily focused on six different issues: (1) the development of specialized objective functions that appropriately represent and summarize the errors between model predictions and observations, (2) the search for efficient optimization algorithms that can reliably solve the hydrologic model calibration problem, (3) the determination of the appropriate quantity and most informative kind of data, (4) the selection of an appropriate numerical solver for the partially structured differential and algebraic equation systems of hydrologic models, (5) the representation of uncertainty, and (6) the development of methods for inferring and refining the mathematical structure and process equations of hydrologic models.

[4] This body of research capitalizes in some way on the "classical" error residual aggregation approach introduced in the early nineteenth century by Adrian Marie Legendre

[1]Department of Civil and Environmental Engineering, University of California, Irvine, California, USA.
[2]Department of Earth System Science, University of California, Irvine, California, USA.
[3]Institute for Biodiversity and Ecosystems Dynamics, University of Amsterdam, Amsterdam, Netherlands.

Corresponding author: J. A. Vrugt, Department of Civil and Environmental Engineering, University of California, Irvine, CA 92697, USA. (jasper@uci.edu)

0043-1397/13/10.1002/wrcr.20354

(1752–1833) and Carl Friedrich Gauss (1777–1855) for fitting simple empirical regression models to noisy data. The least squares method they proposed defines the "best" parameter values as those for which the sum, F_{SLS} of squared residuals,

$$F_{\mathrm{SLS}}(\theta) = \sum_{t=1}^{n} e_t^2(\theta) \tag{3}$$

is at its minimum. In 1822, Gauss was able to state that this approach is optimal if the final residual errors are uncorrelated, with zero mean, and equal variances (homoscedastic). This result is known as the Gauss-Markov theorem.

[5] Several contributions to the hydrologic literature have criticized this historical approach for a lack of treatment of model structural and forcing data errors, which are either assumed "negligibly small" or to be somehow "absorbed" into the error residuals. The residuals are then expected to behave statistically similar as the calibration data measurement error. Yet a posteriori diagnostic checks typically demonstrate that the error residuals exhibit considerable variation in bias, variance, and correlation structures at different parts of the model response. This is in part due to the presence of model structural and forcing (input) data errors whose contribution may, in general, be substantially larger than the (calibration) data measurement error. These errors do not necessarily have any inherent probabilistic properties that can be exploited in the construction of an explicit objective function. While we can assume an (stochastic or deterministic) error model for model structural and forcing data errors, this will be purely for the sake of mathematical convenience.

[6] Some interesting approaches for addressing the limitations of the classical calibration approach, particularly in the context of addressing model parameter and prediction uncertainty, have begun to appear in the literature. These include the limits of acceptability approach [*Beven*, 2006; *Blazkova and Beven*, 2009], the Bayesian Total Error Analysis framework of Kavetski and coworkers [*Kavetski et al*., 2006a, 2006b; *Kuczera et al*., 2006; *Thyer et al*., 2009; *Renard et al*., 2011], the Simultaneous Optimization and Data Assimilation, DiffeRential Evolution Adaptive Metropolis (DREAM), and Particle-DREAM methodologies of Vrugt and coworkers [*Vrugt et al*., 2005, 2012], the stochastic, time-dependent parameter approach of Reichert and coworkers [*Frey et al*., 2011; *Reichert and Mieleitner*, 2009], the generalized likelihood function of *Schoups and Vrugt* [2010] combined with Markov Chain Monte Carlo simulation using DREAM [*Vrugt et al*., 2008a, 2008b, 2009a; *Laloy and Vrugt*, 2012], (Bayesian) model averaging [*Butts et al*., 2004; *Ajami et al*., 2007; *Vrugt and Robinson*, 2007], the hypothetico-inductive data-based mechanistic modeling framework of *Young* [2012], and Bayesian data assimilation [*Bulygina and Gupta*, 2011]. Many of these approaches adopt a Bayesian viewpoint, and relax the assumption of a single "optimum" parameter value in favor of a posterior distribution that accurately recognizes the role of model structural, forcing data, calibration data, and parameter uncertainty (Figure 1).

[7] If we consider the model parameters to be the only source of uncertainty, the posterior parameter distribution, $p(\theta|\tilde{\mathbf{y}}_{1:n})$ can be estimated from Bayes theorem:

$$p(\theta|\tilde{\mathbf{Y}}) = \frac{p(\theta)p(\tilde{\mathbf{Y}}|\theta)}{p(\tilde{\mathbf{Y}})}, \tag{4}$$

where $p(\theta)$ signifies the prior parameter distribution, and $L(\theta|\tilde{\mathbf{Y}}) \equiv p(\tilde{\mathbf{Y}}|\theta)$ denotes the likelihood function. The

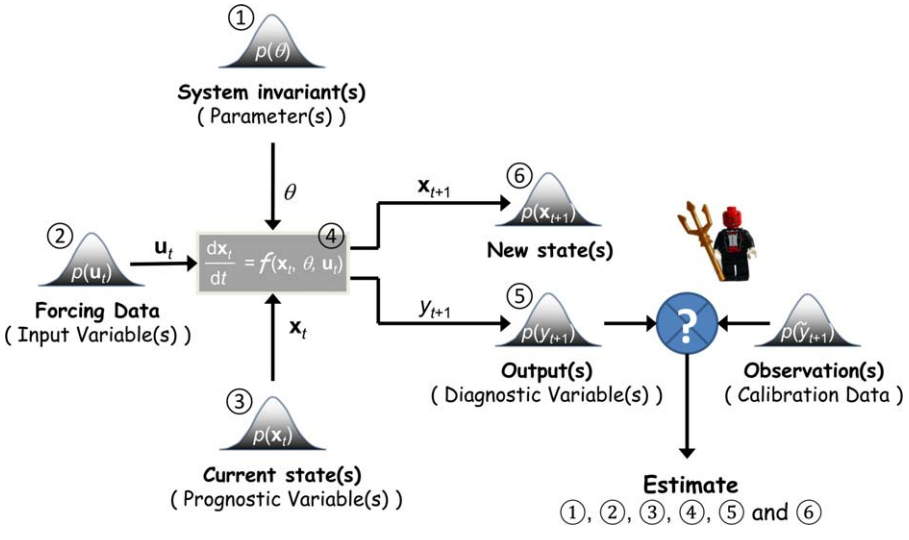

Figure 1. Bayesian model evaluation framework that explicitly recognizes the role of (1) parameter, (2) forcing data, (3) initial state, (4) model structural, (5) output, and (6) state uncertainty. The symbol (circled cross) illuminates the difficulty which likelihood function (and prior distribution/parameterization of the individual error sources) to use to help figure out how the model can be further improved. Approximate Bayesian Computation recognizes the often poor justification of explicit likelihood functions in favor of one or multiple diagnostic metrics that have a better and more compelling diagnostic power than some single average metric of the residual errors.

normalization constant or evidence, $p(\tilde{\mathbf{Y}})$ is required for Bayesian model selection and averaging [*Marshall et al*., 2004], but if our interest is only in the parameters all our statistical inferences (mean, standard deviation, etc.) can be made from the unnormalized density:

$$p(\theta|\tilde{\mathbf{Y}}) \propto p(\theta)L(\theta|\tilde{\mathbf{Y}}). \quad (5)$$

[8] The main culprit now resides in the definition of the likelihood function, $L(\theta|\tilde{\mathbf{Y}})$, that summarizes the overall distance between the model simulations and corresponding observations. If we assume the error residuals to be uncorrelated, Gaussian distributed with constant variance, σ_ν^2, the likelihood function can be written as

$$L(\theta|\tilde{\mathbf{Y}}) = \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\hat{\sigma}_\nu^2}} \exp\left[-\frac{1}{2}\hat{\sigma}_\nu^{-2}(\tilde{y}_t - y_t(\theta))^2\right], \quad (6)$$

where $\hat{\sigma}_\nu$ is an estimate of the standard deviation of the measurement error. The value of $\hat{\sigma}_\nu$ can be specified a priori based on knowledge of the measurements errors, or alternatively its value can be inferred simultaneously with the values of θ [*Vrugt et al*., 2008b; *Laloy and Vrugt*, 2012].

[9] For reasons of algebraic simplicity and numerical stability, it is often convenient to consider the log-likelihood function, $\ell(\theta|\tilde{\mathbf{Y}})$ rather than $L(\theta|\tilde{\mathbf{Y}})$

$$\ell(\theta|\tilde{\mathbf{Y}}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{\sigma}_\nu^2) - \frac{1}{2}\hat{\sigma}_\nu^{-2}\sum_{t=1}^{n}(\tilde{y}_t - y_t(\theta))^2, \quad (7)$$

which is equivalent to

$$\ell(\theta|\tilde{\mathbf{Y}}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{\sigma}_\nu^2) - \frac{1}{2}\hat{\sigma}_\nu^{-2}F_{\mathrm{SLS}}(\theta). \quad (8)$$

[10] This equation elucidates the relationship between traditional least squares fitting and Bayesian inference. This formulation can be extended to weighted least squares, if we assume a heteroscedastic error model, and let $\hat{\sigma}_\nu^2$ depend on the actual observation.

[11] The choice of an adequate likelihood function, $L(\theta|\tilde{\mathbf{Y}})$ has been the subject of considerable debate in the hydrologic and statistical literature. In the past years, various improvements to equation (7) have been proposed to handle nontraditional residual distributions [*Schoups and Vrugt*, 2010; *Smith et al*., 2010]. Despite this progress made, model and input data errors do not necessarily have any inherent probabilistic properties that are easily exploited in the construction of an objective function. Moreover, it has been increasingly recognized that such average measures of model/data similarity inherently lack the power to provide a meaningful comparative evaluation of the consistency in model form and function. The very construction of the likelihood function, as a summary variable of the (usually averaged) properties of the error residuals, dilutes and mixes the available information into an index having little remaining correspondence to specific behaviors of the system. Whereas, this classical approach works for relatively simple (low order) models and allows for some treatment of uncertainty, it is fundamentally weak by design: (a) it fails to exploit the interesting information in the data, and (b) it fails to relate the information to characteristics of the model in a diagnostic manner [*Gupta et al*., 2008].

[12] The limitations of current (Bayesian) strategies for model-data fusion have stimulated *Gupta et al*. [2008] to advocate a diagnostic based approach to model evaluation that has a more clear and compelling diagnostic power to detect structural model deficiencies. This approach uses signature based indices to measure theoretically relevant parts of system behavior, and diagnostic evaluation proceeds with analysis of the behavioral (signature) similarities and differences between the system data and corresponding model simulation. Ideally, these differences are then related to individual process descriptions, and model correction takes place by refining/improving these respective components of the model. *Gupta et al*. [2008] state that (Page 9) "...The diagnostic evaluation approach can be framed within a Bayesian uncertainty framework in a straightforward (though not necessarily trivial) manner." But limited guidance is provided in the paper how to do this in practice. Practical experience suggests that it is not particularly easy to formulate and built a signature based likelihood function with roots in formal statistical theory.

[13] In this rapid communication we introduce approximate Bayesian computation (ABC) as a possible vehicle for diagnostic model evaluation and uncertainty quantification. This yet existing statistical likelihood-free methodology has many ideas in common with a signature based approach, but benefits from a sound theoretical underpinning that will inspire confidence in the computational results. Two simple case studies are presented that illustrate the ABC methodology using a simple least squares fitting problem, and conceptual hydrologic model.

2. Approximate Bayesian Computation

[14] This paper draws inspiration from recent developments in population and evolutionary genetics [*Pritchard et al*., 1999; *Beaumont et al*., 2002], and introduces "likelihood-free" inference to hydrologic modeling and uncertainty quantification. This approach was introduced in the statistical literature about three decades ago [*Diggle and Gratton*, 1984], and is especially useful in situations where evaluation of the likelihood is computationally prohibitive, or for cases when an explicit likelihood (objective) function cannot be justified. This class of methods is also referred to as ABC and is currently a "hot" topic in statistics [*Marjoram et al*., 2003; *Sisson et al*., 2007; *Del Moral et al*., 2011; *Joyce and Marjoram*, 2008; *Grelaud et al*., 2009; *Ratmann et al*., 2009]. The premise behind ABC is that θ^* should be a sample from the posterior distribution as long as the distance between the observed and simulated data, hereafter referred to as $\rho(\tilde{\mathbf{Y}}, \mathbf{Y}(\theta^*))$ is less than some small positive value, ϵ [*Marjoram et al*., 2003; *Sisson et al*., 2007]. Thus, ABC methods bypass the evaluation of the likelihood function and retain the proposal, θ^* if

$$\rho(\tilde{\mathbf{Y}}, \mathbf{Y}(\theta^*)) \leq \epsilon \quad (9)$$

[15] This converges to the true posterior distribution, $p(\theta|\tilde{\mathbf{Y}})$ when $\epsilon \to 0$ [*Pritchard et al*., 1999; *Beaumont*

et al., 2002; *Ratmann et al*., 2009; *Turner and van Zandt*, 2012]. Note that a recent paper by *Nott et al*. [2012] points out the deeper theoretical connection between Generalized likelihood uncertainty estimation (GLUE) and ABC approaches.

[16] A pseudocode of the generic ABC approach is given below.

Algorithm 1: Rejection sampler (ABC-REJ)
for $i = 1, \ldots, N$ **do**
  repeat
    generate θ^* from the prior distribution, $\theta^* \sim p(\theta)$
    simulate $\mathbf{Y}$ using the model, $\mathbf{Y} \leftarrow f(\theta^*|\cdot)$
  until $\rho(\tilde{\mathbf{Y}}, \mathbf{Y}(\theta^*)) \leq \epsilon$
  set $\theta_i \leftarrow \theta^*$
  set $w_i \leftarrow \frac{1}{N}$
end for

[17] In words, the ABC algorithm proceeds as follows. First we sample a candidate point, θ^* from some prior distribution, $p(\theta)$. We then use this proposal to simulate the output of the model, $\mathbf{Y} \leftarrow f(\theta^*|\cdot)$. We then compare the simulated data, $\mathbf{Y}$ to the observed data $\tilde{\mathbf{Y}}$ using a distance function, $\rho(\tilde{\mathbf{Y}}, \mathbf{Y}(\theta^*))$. If this distance function is smaller than some predefined tolerance value, ϵ then the simulation is close enough to the observations that the candidate point, θ^* has some nonzero probability of being in the approximate posterior distribution, $\hat{p}(\theta|\rho(\tilde{\mathbf{Y}}, \mathbf{Y}) \leq \epsilon)$. Note that for very small values of ε, the rejection rates can be dramatically high, and Algorithm 1 thus very inefficient.

[18] For sufficiently complex models and large data sets the probability of happening upon a simulation run that yields precisely the same simulation as the calibration data set will be very small, often unacceptably so. To resolve this problem, it is often convenient to define $\rho(\tilde{\mathbf{Y}}, \mathbf{Y}(\theta^*))$ as a distance between summary statistics, $S(\mathbf{Y}(\theta^*))$ and $S(\tilde{\mathbf{Y}})$ of the simulated and observed data, respectively. If the distance between those summary statistics, $\rho(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\theta^*)))$ is smaller than ε the sample is retained. For example, $S(\cdot)$ could be the sample mean, $\rho(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\theta^*))) = |\text{mean}(\tilde{\mathbf{Y}}) - \text{mean}(\mathbf{Y}(\theta^*))|$. Ideally, the chosen summary statistic, $S(\cdot)$ is sufficient for θ and thus provides as much information for the parameters as the original data set itself. If the exact (perfect) likelihood function is unknown, it will be difficult to determine a sufficient statistic. In practice, one could use multiple different summary statistics that each capture different aspects/signatures/patterns of the input-output response [*Ratmann et al*., 2009]. Arguably, this approach has many elements in common with multicriteria optimization approaches, yet benefits from a proper statistical foundation and only investigates the compromised region of the Pareto solution space.

[19] The user is free to select which summary statistics to use. But they should be chosen carefully, so as to extract meaningful information from the available data and ensure adequate posterior convergence. Perhaps more important from the viewpoint of the present paper, the ABC methodology is particularly useful for diagnostic model calibration and evaluation. By defining summary metrics so that they each measure different parts of system behavior, the posterior distribution could help to pinpoint which individual components of the model are in need of improvement. The second case study will help to illustrate this in more detail.

3. Case Study I: Toy Problem

[20] To illustrate the ABC methodology, consider a simple linear regression model, $y_t = \theta_1 t + \theta_2$ as a toy problem. A synthetic data set of observations, $\tilde{\mathbf{Y}} = \{\tilde{y}_1, \ldots, \tilde{y}_n\}$ was created for $n = 100$ equidistant points between $t = 0$ and $t = 10$ using $\theta_1 = \frac{1}{2}$ and $\theta_2 = 5$. These observations were subsequently perturbed with white noise, $\tilde{\mathbf{Y}} \leftarrow N(\tilde{\mathbf{Y}}, \sigma_\nu^2)$ with $\sigma_\nu = 0.25$ to represent the effect of data measurement errors. Figure 2 plots the synthetic data (solid circles), and the original simulation of the linear regression model (solid black line).

[21] Figure 3 (top) presents histograms of the posterior marginal distributions of the parameters θ_1 and θ_2 derived from least squares fitting (analytical solution). The asterisks indicate the true parameter values used to create the synthetic data. As expected, the marginal distributions are well defined, with modes of the slope and intercept that coincide with their actual values used to create the observations.

[22] We now use ABC with summary statistic the mean of the actual data, $\rho(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\theta))) = |\text{mean}(\tilde{\mathbf{Y}}) - \text{mean}(\mathbf{Y}(\theta))|$. A uniform prior with $\theta_1 \in [0, 1]$ and $\theta_2 \in [-10, 10]$ was used in all our calculations. To increase computational efficiency, we used the ABC population Monte Carlo (PMC) scheme of *Turner and van Zandt* [2012], the details of which are given in Appendix A. In short, this ABC-PMC sampler starts out as rejection sampler (ABC-REJ) during the first iteration, $j = 1$, but using a much larger initial value for ε. During each successive next step, $j = \{2, \ldots, J\}$ the value of ε is iteratively decreased and

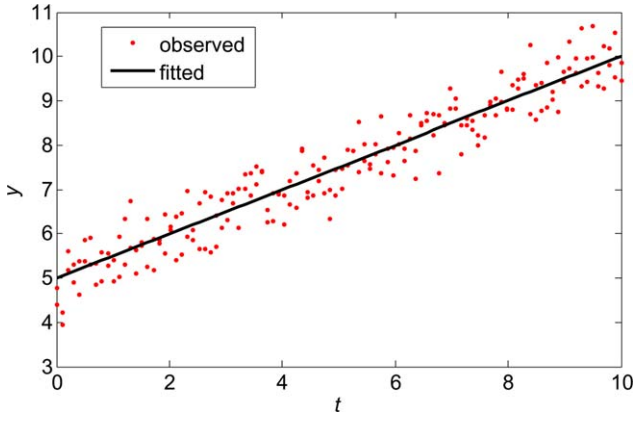

Figure 2. The original regression model, $\mathbf{Y} \leftarrow \frac{1}{2}t + 5, t \in \{0, 10\}$ (solid black line) and perturbed data (solid circles) that was used to illustrate the ABC methodology. Summary statistics of the measured and simulated data are used to derive estimates of the slope and intercept of the regression function. This type of inference approach differs fundamentally from Bayesian and Frequentist parameter estimation methods that work on the error residuals rather than some summary statistics of the observed and simulated data to find the desired parameter values.

the proposal distribution, $q_j\left(\theta_k^{j-1},\cdot\right) = N_d\left(\theta_k^{j-1},\Sigma^j\right)_{(j>1)}$ adapted using $\Sigma^j = \mathrm{Cov}\left(\theta_1^{j-1},\ldots,\theta_N^{j-1}\right)$ with θ_k drawn from a discrete multinomial distribution, $\Im\left(\theta_{1:N}^{j-1}|\mathbf{w}_{1:N}^{j-1}\right)$ where $\mathbf{w}_{1:N}^{j-1}$ denote the posterior weights $\left(w_l^{j-1}\geq 0;\sum_{l=1}^{N} w_l^{j-1}=1\right)$. Through a sequence of successive (multi)normal proposal distributions the prior sample is thus iteratively refined until a sample of the posterior distribution is obtained. This approach, similar in spirit as the adaptive Metropolis sampler of *Haario et al*. [1999, 2001] receives a much higher sampling efficiency than ABC-REJ, particularly for cases where the prior sampling distribution, $p(\theta)$ is a poor approximation of the actual posterior distribution.

[23] The results of ABC-PMC are presented in the second plot (from top) in Figure 3 and correspond to $\epsilon = 0.005$. Smaller values of ε give similar results as will be demonstrated later. The marginal distributions settle around their "true" values, but demonstrate significant dispersion, much larger than what can be expected from the data. From the results it should be evident that the mean of the observations is a relatively weak (insufficient) statistic with poor diagnostic power. Indeed, many combinations of the slope and intercept exist with mean ($\mathbf{Y}(\theta)$) similar to that of the observations. We therefore add, as second metric, the standard deviation of the data (observations), std $\left(\tilde{\mathbf{Y}}\right)$ using

$$\rho\left(S\left(\tilde{\mathbf{Y}}\right),S(\mathbf{Y}(\theta^*))\right) = \max\left(\left|\mathrm{mean}\left(\tilde{\mathbf{Y}}\right)-\mathrm{mean}\left(\mathbf{Y}(\theta^*)\right)\right|,\ \left|\mathrm{std}\left(\tilde{\mathbf{Y}}\right)-\mathrm{std}\left(\mathbf{Y}(\theta^*)\right)\right|\right), \quad (10)$$

as a distance function. Figure 3 (bottom) shows the corresponding marginal posterior parameter distributions of θ_1 and θ_2. The histograms now show a striking resemblance with those obtained for the least squares regression analysis. With two sufficient diagnostics we derive virtually similar posterior histograms as those from a classical first-order statistical analysis.

[24] We now analyze the effect of the choice of ε on the ABC results. Figure 4 plots the ABC-PMC derived posterior standard deviation of the slope (red line) and intercept (blue line) of the linear model for $J = 11$ successively smaller values of ε. The standard deviation provides a reasonable diagnostic of the dispersion (width) of the posterior sample, and can be used to help judge when convergence of ABC-PMC to a limiting distribution has been achieved. The width of the marginal posterior distributions of θ_1 and θ_2 rapidly decreases during the first few iterations, but then settles around a stable value (slope = 0.014 and intercept = 0.079) for $\epsilon \leq 0.05$. These numerical results confirm that in the limit of $\epsilon \to 0$ the ABC-PMC sampler converges to the exact posterior distribution. For the present analysis $\epsilon = 0.05$ seems adequate. Lower values of ε provide similar posterior estimates, yet unnecessarily increase the computational burden of the ABC analysis.

4. Case Study II: Conceptual Hydrologic Modeling

[25] We now move on to a more reasonable example, and apply the ABC methodology to rainfall-runoff modeling. We use daily data of mean areal precipitation, $\tilde{\mathbf{P}} = \{\tilde{p}_1,\ldots,\tilde{p}_n\}$, mean areal potential evaporation, and streamflow, $\tilde{\mathbf{Y}} = \{\tilde{y}_1,\ldots,\tilde{y}_n\}$ from the French Broad River basin at Asheville, North Carolina. This is the wettest watershed of the original 12 MOPEX basins described in the study by *Duan et al*. [2006]. We use a lumped conceptual hydrologic model, described in detail by *Schoups and Vrugt* [2010], and part of the DREAM software package. The model transforms rainfall into runoff at the watershed

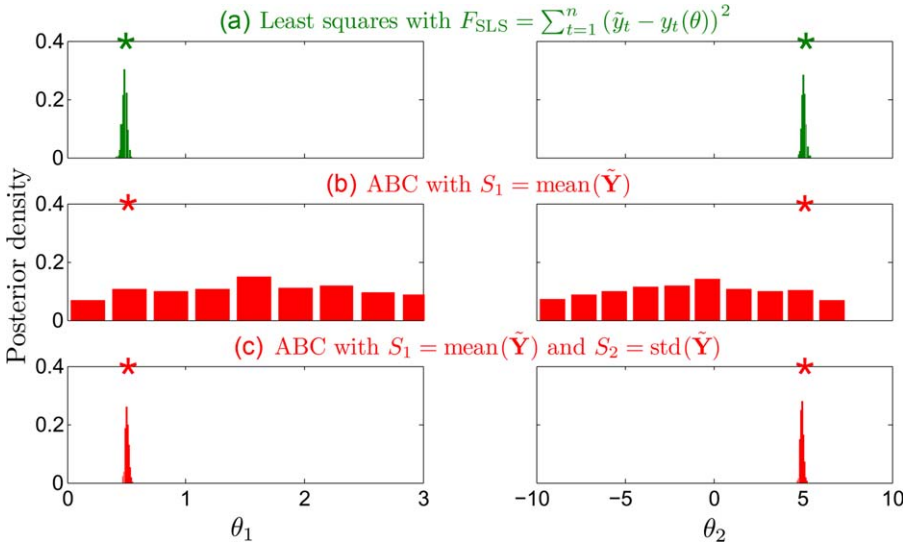

Figure 3. Marginal posterior distribution of the slope, θ_1 and intercept, θ_2 derived from (top) least squares fitting and ABC inference using (middle) one (mean of data) and (bottom) two (mean and standard deviation of data) summary diagnostic(s). The asterisks in each plot denote the actual parameter values used to create the synthetic data, $\hat{\mathbf{Y}}$.

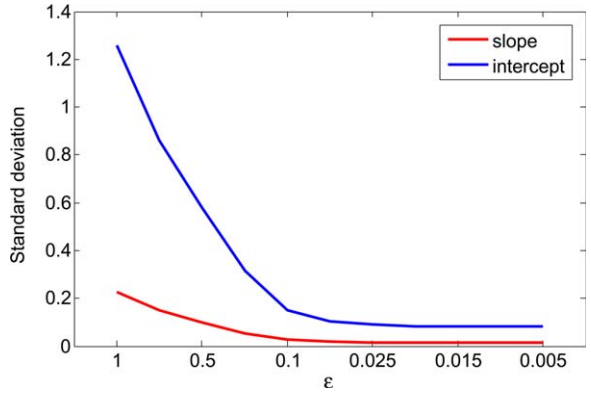

Figure 4. Relationship between the value of ε used in ABC-PMC and the posterior standard deviation of the slope (red line) and intercept (blue line). Results confirm theory that if $\epsilon \rightarrow 0$ the posterior statistics converge to their target values. A value of $\epsilon = 0.05$ generates stable results.

outlet using explicit process descriptions of interception, throughfall, evaporation, runoff generation, percolation, and surface and subsurface routing. Table 1 summarizes the seven different model parameters, and their prior uncertainty ranges.

[26] Figure 5 presents the results of the ABC analysis using the mean and standard deviation of the discharge data as summary statistics. Figure 5a presents time series plots of the 95% posterior prediction uncertainty ranges (due to parameter uncertainty) and corresponding observations (solid circles). Figures 5b–5f plot histograms of the marginal posterior distribution of four (randomly) selected model parameters, and the posterior Root Mean Square Error (RMSE) values. The streamflow prediction uncertainty ranges generally track the observed data and encapsulate about 88% of the observations. This is a rather encouraging result, and convincingly demonstrates the utility of summary statistics for model calibration purposes. Nevertheless, the parameters are poorly defined with posterior ranges that extend a large part of the prior ranges. This result is perhaps not surprising. The mean and standard deviation of the discharge observations are relatively weak summary metrics to judge the "distance" between the model simulation and data, and thus exhibit insufficient diagnostic power to adequately constrain the model parameter space. Note that the posterior RMSE values exhibit a log-normal distribution with values ranging between 0.67 and 2.04 mm/d. These values and their range is much larger than what can be expected from a traditional explicit likelihood function that acts directly on the error residuals (more about this later).

Table 1. Prior Uncertainty Ranges of Hydrologic Model Parameters

Parameter	Symbol	Minimum	Maximum	Units
Aximum interception	$I_{\max}$	0	10	mm
Soil water storage capacity	$S_{\max}$	10	1000	mm
Maximum percolation rate	$Q_{\max}$	0	100	mm/d
Evaporation parameter	α_E	0	100	
Runoff parameter	α_F	−10	10	
Time constant, fast reservoir	K_F	0	10	days
Time constant, slow reservoir	K_S	0	150	days

[27] We repeat the analysis, but now with summary metrics that capture hydrologically relevant aspects of the watersheds behavior. We consider three such hydrologic signatures including the annual runoff coefficient, the annual base flow index [*Gupta et al*., 2008; *Yilmaz et al*., 2008], and the flow duration curve (FDC) [*Searcy*, 1959; *Westerberg et al*., 2011]. The annual runoff coefficient, S_1 and annual base flow index, S_2 are respectively defined as

$$S_1(\tilde{\mathbf{Y}}) = \sum_{t=1}^{365} \frac{\tilde{y}_t}{\tilde{p}_t} \quad S_2(\tilde{\mathbf{Y}}) = \sum_{t=1}^{365} \frac{\tilde{y}_{b,t}}{\tilde{y}_t}, \tag{11}$$

where t signifies the day in the hydrologic year (1 October to 30 September), and the base flow contribution, $\tilde{y}_{b,t} \leq \tilde{y}_t$ at time $t = \{1, \ldots, n\}$ is computed using a low-pass filter,

$$\tilde{y}_{b,t} = \phi \tilde{y}_{b,t-1} + \frac{1}{2}(1-\phi)(\tilde{y}_t + \tilde{y}_{t-1}), \tag{12}$$

with ϕ set at 0.925 [*Arnold and Allen*., 1999; *Eckhardt*, 2005; *Carrillo et al*., 2011].

[28] Now we have defined an explicit measure of the annual runoff and base flow index, we are now left with the definition of a summary metric for the FDC. Unfortunately, the nonlinear, hyperbolic, shape of this curve renders a single metric insufficient to adequately capture and summarize the distribution of the flow levels. This follows directly from our previous study that showed that at least two pieces of information (metrics) are needed to adequately constrain the slope and intercept of a linear regression function (see toy problem). To mimic the observed FDC as closely and consistently as possible we use the following function

$$F_t \leftarrow \left[1 + (S_3 \tilde{y}_t)^{S_4}\right]^{\left(\frac{1}{S_4}-1\right)}, \tag{13}$$

where F_t denotes the exceedance probability of the t^{th} streamflow observation, and S_3 (d/mm) and S_4(-) are calibration coefficients that act as our summary metrics to define the functional shape of the curve. This leaves us with four summary statistics for three different hydrologic signatures.

[29] We use the following default distance function in our ABC-PMC sampler

$$\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\theta^*))\big) = \max\big(|S_k(\tilde{\mathbf{Y}}) - S_k(\mathbf{Y}(\theta^*))|; k = \{1, \ldots, 4\}\big), \tag{14}$$

to help decide whether to accept θ^* or not. Model parameters that simultaneously mimic each of the measured signatures are expected to represent system properties and accurately represent the watersheds hydrologic response to rainfall. In our calculations with ABC-PMC we use $\varepsilon = \{1.0, 0.75, 0.5, 0.25, 0.1, 0.05, 0.025, 0.02, 0.01\}$ (see Appendix A) which led to an acceptance rate of about 1%. In

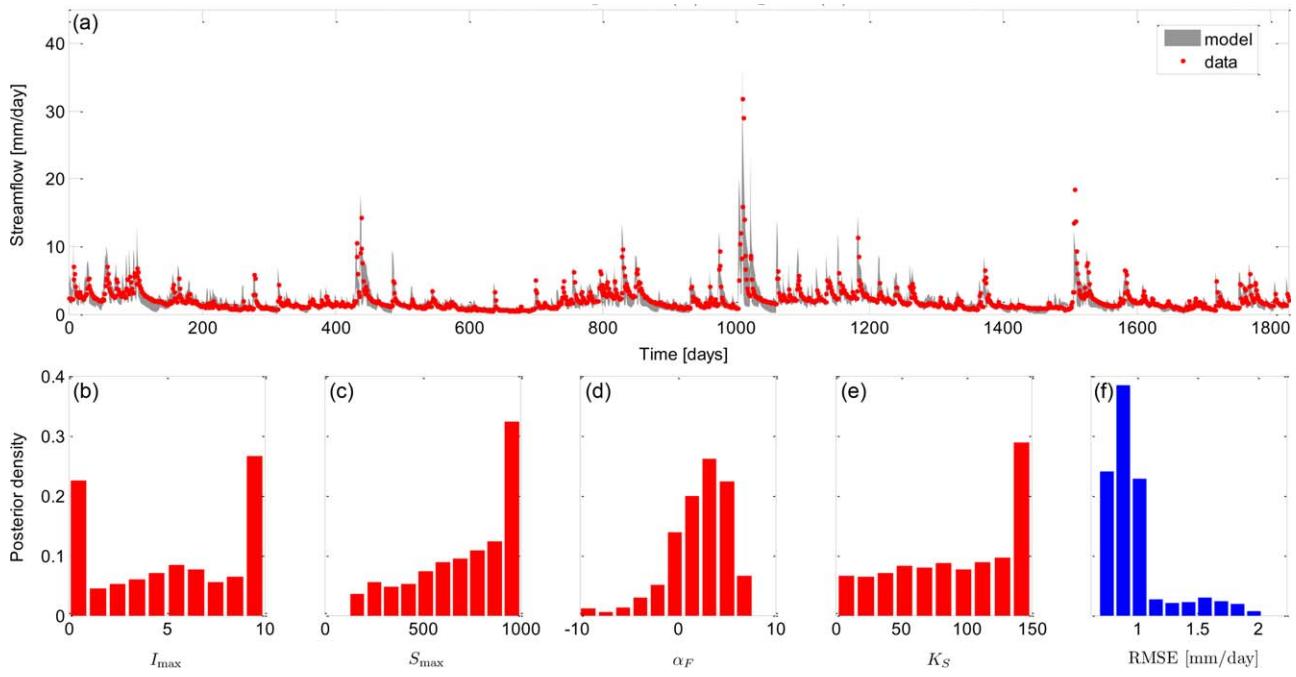

Figure 5. ABC inference for the French Broad River basin using the mean and standard deviation of the discharge data as summary statistics: (a) 95% posterior prediction uncertainty ranges (gray region) and corresponding discharge observations (solid circles), (b–e) histograms of the marginal posterior parameter distributions of (b) I_{max}, (c) S_{max}, (d) α_F, and (e) K_S, and (f) histogram of posterior RMSE values. The posterior uncertainty ranges nicely encapsulate the actual data, but with parameter distributions that appear too dispersed.

another recent paper [*Sadegh and Vrugt*, 2013], we have presented an alternative variant of PMC that adaptively selects the sequence of ε values based on the different generations of $\rho(\cdot)$ values of the accepted samples.

[30] Note that equation (13) is equivalent to the water retention model of *van Genuchten* [1980] but with $\theta_s = 1$ and $\theta_r = 0$ so that $0 \leq F_t \leq 1$. A subsequent paper will demonstrate the ability of equation (13) to closely represent the FDC of a large range of watersheds with widely varying hydrologic responses. Here we limit our attention to the French Broad River basin, and plot in Figure 6 the observed (red dots) and model predicted (solid back line) FDC. The least squares fit of the proposed two-parameter model can be considered excellent with a RMSE of about 0.01 mm/d and optimized values of $S_3 = 0.74$ and $S_4 = 3.34$.

[31] Figure 7 presents the results of ABC using hydrologic summary statistics, $S_1 \rightarrow S_4$. The posterior parameter ranges (Figures 7b–7e) have reduced somewhat compared to Figure 6, but continue to extent a large portion of the prior distribution. A wide range of model parameterizations can be found that honor the observed annual runoff coefficient, annual base flow index, and distribution of flow levels but with simulated streamflow response that does not always match the observed hydrographs (Figure 7a). In particular, the recession periods appear to be well resolved by the model but the peak flows are almost systematically underestimated. This deviation appears systematic and is thus unlikely to be explained by errors in the discharge and precipitation data only.

[32] Nothing prevents us from using additional signatures in the ABC analysis that better capture the functional form of the rising limb of the hydrograph. What comes to mind are the average magnitude of peak flow and time to peak. Preliminary ABC runs illustrate however that the model is unable to simultaneously satisfy all these

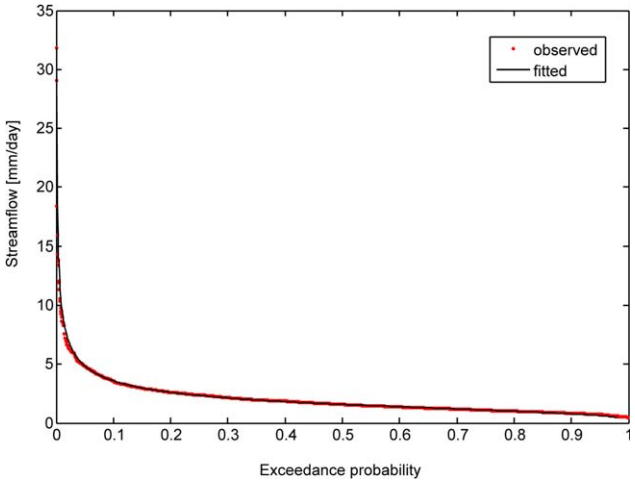

Figure 6. Observed (red dots) and model fitted (solid black line) FDC of the French Broad River basin at Asheville, North Carolina, USA. The two-parameter model closely matches the observed FDC with prediction error on the order of 0.01 mm/d. The optimized parameters, S_3 and S_4 of the FDC model are used as hydrologic signatures in our ABC analysis.

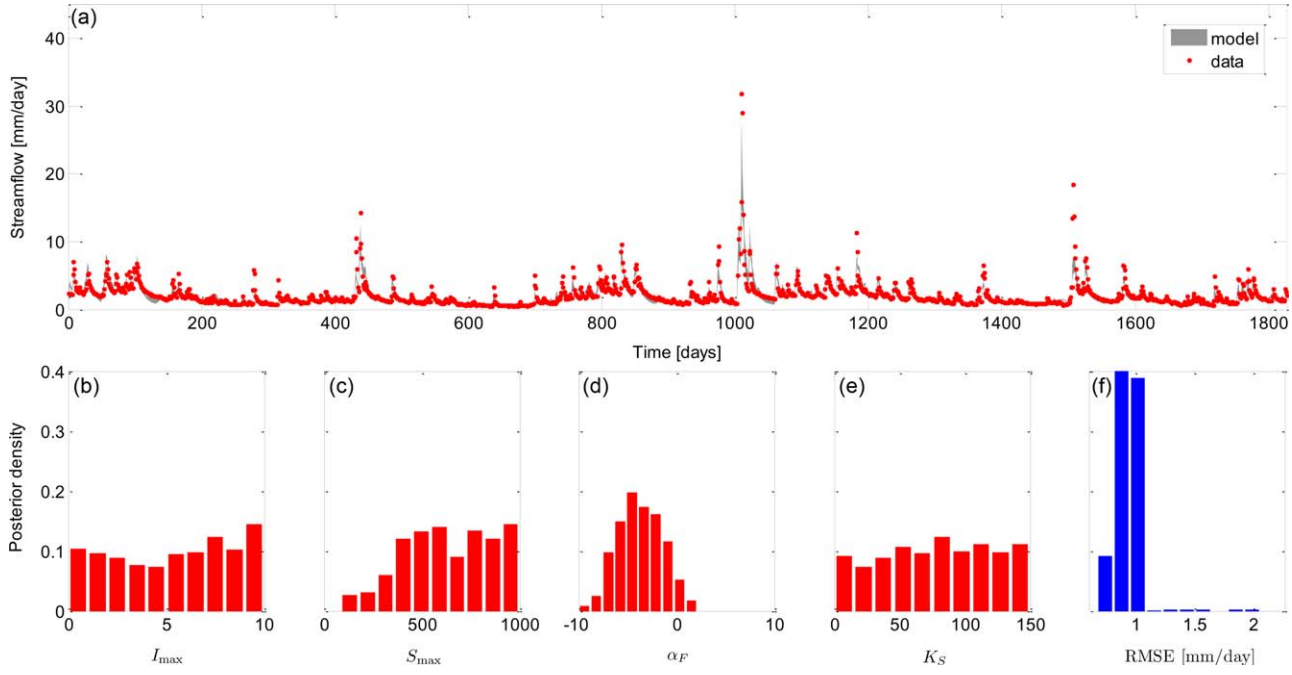

Figure 7. ABC inference for the French Broad River basin using summary metrics $S_1 \rightarrow S_4$ that represent the observed annual runoff coefficient, annual base flow index, and flow duration curve: (a) 95% posterior prediction uncertainty ranges (gray region) and corresponding discharge observations (solid circles), (b–e) histograms of the marginal posterior parameter distributions of (b) $I_{\max}$, (c) $Q_{\max}$, (d) α_F, and (e) K_S, and (f) histogram of posterior RMSE values. The posterior parameter ranges cover a large portion of their prior uncertainty ranges. The modeled discharge time series satisfy the three hydrologic signatures used in the analysis, but deviate at least periodically significantly from the observed watersheds response to rainfall. The fit to streamflow data can be improved significantly if additional metrics are incorporated in the ABC analysis that more directly honor the functional shape of the driven and nondriven part of the hydrograph.

hydrologic signatures. This points to a deficiency in the model structure, that prevents the model from being able to simultaneously fit both the peak flow and nondriven part of the hydrograph. By stepwise increasing the number of hydrologic signatures, we can exactly pinpoint which part (process description) of the model is deficient and in need of further improvement/refinement. In this particular case it is obvious that some level of model malfunction occurs in the driven part of the hydrograph, likely related to an erroneous process description of surface runoff and/or presence of preferential flow that is not accounted for by the model.

[33] Note that the classical model calibration approach would not provide such guidance. For instance, consider the (least squares) likelihood function presented in equation (8) that minimizes in some average mathematical sense the size of the error residuals. This approach indeed gives substantially lower RMSE values (0.56–0.59 mm/d) than the ABC-signature based analysis (Figure 7f) (0.67–2 mm/d and median value of about 0.93) but introduces erroneous parameter values that fit the driven part of the hydrograph at the expense of the recession periods [see *Schoups and Vrugt*, 2010, Figure 6]. The parameters are thus adjusted in such a way that they compensate (optimally) for the deficiencies in the model structure but with simulated signatures that deviate considerably from their observed counterparts. This not only severely diminishes our chances to detect when and where the model is in error, but also complicates finding useful regionalization relationships.

[34] Figure 8 presents scatter plots of the $N = 1{,}000$ posterior samples derived with the ABC-PMC algorithm for six selected parameter pairs. It is evident that the adaptive capability of the ABC-PMC algorithm enables it to track the strongly nonlinear $p(\theta|\tilde{\mathbf{Y}})$ surface. Whereas the histograms previously depicted in Figures 7b–7e suggest that the model parameters are poorly identifiable by calibration against the annual runoff coefficient, annual base flow index, and FDC, bivariate plots of the posterior samples show that the behavioral solutions occupy only a relatively small part of the prior parameter space. The strongly nonlinear, banana-shaped, correlation of the α_F– K_F samples is particularly interesting, and diagnoses a structural problem with the surface runoff and fast flow component of the model. This confirms our earlier conclusion that our process description of the driven part of the hydrograph is (somewhat) deficient and in need of improvement. This should improve the ability of the model to fit peak flow.

[35] This concludes our numerical simulations. A subsequent paper will investigate in more detail the reasons for this model malfunctioning and proceeds with correction. We now conclude this paper with a short discussion.

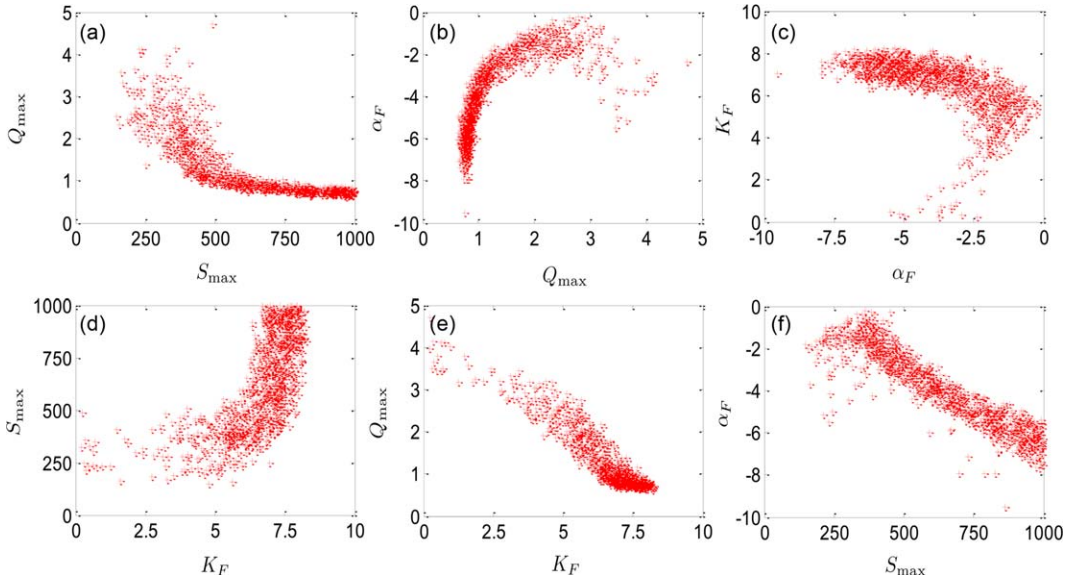

Figure 8. Bivariate plots of the ABC derived posterior samples of six different parameter pairs for the French Broad River basin using summary metrics $S_1 \rightarrow S_4$: (a) $S_{\max} - Q_{\max}$, (b) $Q_{\max} - \alpha_F$, (c) $\alpha_F - K_F$, (d) $K_F - S_{\max}$, (e) $K_F - Q_{\max}$, and (f) $S_{\max} - \alpha_F$. The individual parameters appear poorly identified, but the behavioral parameter space encompasses a well-defined region interior to the prior parameter space.

5. Discussion

[36] From the examples presented we see that the ABC framework has many desirable characteristics that lend itself useful for diagnostic model evaluation. With the results of this paper in mind, three interesting ideas that immediately follow include

[37] 1. The impact of precipitation data uncertainty on the calibrated parameter estimates can be drastically reduced by developing hydrologic signatures that are insensitive to errors in the hyetograph. This would not only significantly enhance the prospects of detecting model structural errors, but also provide new perspectives for regionalization studies. We also posit that such procedure would drastically reduce the required length of the data to derive stable parameter estimates [e.g., *Schoups and Vrugt* 2010]. Two related shape descriptors, the rising and declining limb density [*Shamir et al*., 2005] could prove particularly useful in this context, and are easily integrated in the ABC framework presented herein.

[38] 2. There appears to be a close (theoretical) connection between GLUE and ABC. This has recently also been elucidated by *Nott et al*. [2012]. The limits of acceptability approach of GLUE [*Beven*, 2006; *Blazkova and Beven*, 2009] is a special variant of ABC in that each individual observation is used as summary statistic [*Sadegh and Vrugt*, 2013]. This confirms our earlier conclusions reported in *Vrugt et al*. [2008b] that formal and informal Bayesian methods might have more common ground than the literature and ongoing debates might seem to suggest.

[39] 3. The ABC methodology provides a more rigorous statistical footing for multicriteria methods that use different parts of the model response or multiple different data types for calibration purposes. This is a plausible extension of the work of *Reichert and Schuwirth* [2012] who framed the use of multiple calibration objectives in a formal probabilistic (Bayesian) framework.

[40] Future papers by the authors on these topics are forthcoming.

6. Summary and Conclusions

[41] The chronic historical deficit of robust and powerful model evaluation methods for systematically engaging complex environmental models with field data is increasingly limiting our ability to diagnose, detect, and resolve model structural deficiencies. Diagnostic model evaluation has been proposed to illuminate to what degree a representation of the real world has been adequately achieved and how the model should be improved.

[42] In this paper, we introduced ABC as an existing statistical vehicle for diagnostic model evaluation. This method readily incorporates signature based metrics, and has its roots within a proper statistical context. Two different case studies involving a simple linear toy problem, and conceptual hydrologic model were used to introduce the ABC framework as paradigm for diagnostic model evaluation. Results demonstrate that ABC is relatively easy to implement, and readily employs signature based indices to analyze and pinpoint which part of the model is malfunctioning and in need of further improvement.

Appendix A

[43] Suppose some measurement data $\tilde{\mathbf{Y}} = \{\tilde{y}_1, \ldots, \tilde{y}_n\}$, and a model that predicts $\mathbf{Y} \leftarrow f(\theta|\cdot)$ with parameter values, $\theta \in \Theta \in \mathbb{R}^d$. We define a prior distribution, $p(\theta)$ and a vector with decreasing tolerance values, $\varepsilon = \{\epsilon_1, \ldots, \epsilon_J\}$ so that $\epsilon_{j+1} < \epsilon_j, \forall j \in \{2, \ldots, J\}$. The ABC population Monte Carlo method proceeds as follows [*Turner and van Zandt*, 2012]

Algorithm 2: ABC population Monte Carlo sampler

At iteration $j = 1$,
for $i = 1, \ldots, N$ **do**
  while $\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y})\big) > \epsilon_1$ **do**
    Sample θ^* from the prior, $\theta^* \sim p(\theta)$
    Simulate data $\mathbf{Y}$ using $\theta^*, \mathbf{Y} \leftarrow f(\theta^*|\cdot)$
    Calculate $\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\theta^*))\big)$
  end while
  Set $\theta_i^1 \leftarrow \theta^*$
  Set $w_i^1 \leftarrow \frac{1}{N}$
end for
Set $\sigma_1^2 \leftarrow 2\mathrm{Cov}\left(\theta_{1:N}^1\right)$,
At iteration $j > 1$,
for $j = 2, \ldots, J$ **do**
  for $i = 1, \ldots, N$ **do**
    while $\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y})\big) > \epsilon_j$ **do**
      Sample θ^* from the previous iteration, $\theta^* \sim \theta_{1:N}^{j-1}$ with probability $\mathbf{w}_{1:N}^{j-1}$
      Perturb θ^* by sampling $\theta^{**} \sim N(\theta^*, \sigma_{j-1}^2)$
      Simulate data $\mathbf{Y}$ using $\theta^{**}, \mathbf{Y} \leftarrow f(\theta^{**}|\cdot)$
      Calculate $\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\theta^{**}))\big)$
  end while
  Set $\theta_i^j \leftarrow \theta^{**}$
  Set $w_i^j \leftarrow \dfrac{p\left(\theta_i^j\right)}{\sum_{k=1}^{N} w_k^{j-1} q\left(\theta_k^{j-1}|\theta_i^j, \sigma_{j-1}^2\right)}$
  end for
  Set $\sigma_j^2 \leftarrow 2\mathrm{Cov}\left(\theta_{1:N}^j\right)$
end for

[44] The updating of the weights requires evaluation of a multivariate normal pdf, $q(\cdot|b, c)$ with mean b and variance c. This concludes the pseudocode of the population Monte Carlo sampler.

[45] **Acknowledgments.** Both authors highly appreciate the support and funding from the UC-Lab Fees Research Program Award 237285. The comments of the three anonymous referees have improved the current version of this manuscript. The ABC method used herein is available upon request from the first author (jasper@uci.edu).

References

Ajami, N. K., Q. Duan, and S. Sorooshian (2007), An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction, *Water Resour. Res.*, *43*, W01403, doi:10.1029/2005WR004745.

Arnold, J. G., and P. M. Allen (1999), Automated methods for estimating baseflow and ground water recharge from streamflow records, *J. Am. Water Resour. Assoc.*, *35*, 411–424.

Beaumont, M. A., W. Zhang, and D. J. Balding (2002), Approximate Bayesian computation in population genetics, *Genetics*, *162*(4), 2025–2035.

Beven, K. (2006), A manifesto for the equifinality thesis, *J. Hydrol.*, *320*, 18–36, doi:10.1016/j.jhydrol.2005.07.007.

Blazkova, S., and K. Beven (2009), A limits of acceptability approach to model evaluation and uncertainty estimation in flood frequency estimation by continuous simulation: Skalka catchment, Czech Republic, *Water Resour. Res.*, *45*, W00B16, doi:10.1029/2007WR006726.

Bulygina, N., and H. Gupta (2011), Correcting the mathematical structure of a hydrological model via Bayesian data assimilation, *Water Resour. Res.*, *47*, W05514, doi:10.1029/2010WR009614.

Butts, M. B., J. T. Payne, M. Kristensen, and H. Madsen (2004), An evaluation of the impact of model structure on hydrological modeling uncertainty for streamflow simulation, *J. Hydrol.*, *298*, 242–266, doi:10.1016/j.jhydrol.2004.03.042.

Carrillo, G., P. A. Troch, M. Sivapalan, T. Wagener, C. Harman, and K. Sawicz (2011), Catchment classification: Hydrological analysis of catchment behavior through process-based modeling along a climate gradient, *Hydrol. Earth Syst. Sci.*, *15*, 3411–3430, doi:10.5194/hess-15-3411-2011.

Del Moral, P., A. Doucet, and A. Jasra (2011), An adaptive sequential Monte Carlo method for approximate Bayesian computation, *Statistics and Computing*, *2*(5), pp 1009–1020, doi:10.1007/s11222-011-9271-y.

Diggle, P. J., and R. J. Gratton (1984), Monte Carlo methods of inference for implicit statistical models, *J. R. Stat. Soc., Ser. B*, *46*, 193–227.

Duan, Q., et al. (2006), Model parameter estimation experiment (MOPEX): An overview of science strategy and major results from the second and third workshops, *J. Hydrol.*, *320*, 3–17.

Eckhardt, K. (2005), How to construct recursive digital filters for baseflow separation, *Hydrol. Processes*, *19*, 507–515.

Frey, M. P., C. Stamm, M. K. Schneider, and P. Reichert (2011), Using discharge data to reduce structural deficits in a hydrological model with a Bayesian inference approach and the implications for the prediction of critical source areas, *Water Resour. Res.*, *47*, W12529, doi:10.1029/2010WR009993.

Grelaud, A., C. Robert, J. Marin, F. Rodolphe, and J. Taly (2009), ABC likelihood-free methods for model choice in Gibbs random fields, *Bayesian Anal.*, *4*(2), 317–336.

Gupta, H. V., T. Wagener, and Y. Liu (2008), Reconciling theory with observations: Elements of a diagnostic approach to model evaluation, *Hydrol. Processes*, *22*(18), 3802–3813.

Haario, H., E. Saksman, and J. Tamminen (1999), Adaptive proposal distribution for random walk Metropolis algorithm, *Comput. Stat.*, *14*(3), 375–395.

Haario, H., E. Saksman, and J. Tamminen (2001), An adaptive Metropolis algorithm, *Bernoulli*, *7*, 223–242.

Joyce, P., and P. Marjoram (2008), Approximately sufficient statistics and Bayesian computation, *Stat. Appl. Genetics Mol. Biol.*, *7*(1), 1544–6115, doi:10.2202/1544–6115.1389.

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

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

Kuczera, G., D. Kavetski, S. Franks, and M. Thyer (2006), Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterizing model error using storm-dependent parameters, *J. Hydrol.*, *331*, 161–177, doi:10.1016/j.jhydrol.2006.05.010.

Laloy, E., and J. A. Vrugt (2012), High-dimensional posterior exploration of hydrologic models using multiple-try DREAM (*ZS*) and high-performance computing, *Water Resour. Res.*, *48*, W01526, doi:10.1029/2011WR010608.

Marjoram, P., J. Molitor, V. Plagnol, and S. Tavaré (2003), Markov chain Monte Carlo without likelihoods, *Proc. Natl. Acad. Sci. U. S. A.*, *100*(26), 15,324–15,328.

Marshall, L., D. Nott, and A. Sharma (2004), A comparative study of Markov chain Monte Carlo methods for conceptual rainfall-runoff modeling, *Water Resour. Res.*, *40*, W02501, doi:10.1029/2003WR002378.

Nott, D. J., L. Marshall, and J. Brown (2012), Generalized likelihood uncertainty estimation (GLUE) and approximate Bayesian computation: What's the connection?, *Water Resour. Res.*, *48*, W12602, doi:10.1029/2011WR011128.

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun, and M. T. Feldman (1999), Population growth of human Y chromosomes: A study of Y chromosome microsatellites, *Mol. Biol. Evol.*, *16*(12), 1791–1798.

Ratmann, O., C. Andrieu, C. Wiuf, and S. Richardson (2009), Model criticism based on likelihood-free inference, with an application to protein network evolution, *Proc. Natl. Acad. Sci. U. S. A.*, *106*, 1–6.

Reichert, P., and J. Mieleitner (2009), Analyzing input and structural uncertainty of nonlinear dynamic models with stochastic, time-dependent parameters, *Water Resour. Res.*, *45*, W10402, doi:10.1029/2009WR007814.

Reichert, P., and N. Schuwirth (2012), Linking statistical bias description to multiobjective model calibration, *Water Resour. Res.*, *48*, W09543, doi:10.1029/2011WR011391.

Renard, B., D. Kavetski, E. Leblois, M. Thyer, G. Kuczera, and S. W. Franks (2011), Toward a reliable decomposition of predictive uncertainty in hydrological modeling: Characterizing rainfall errors using conditional simulation, *Water Resour. Res.*, *47*, W11516, doi:10.1029/2011WR010643.

Sadegh, M., and J. A. Vrugt (2013), Approximate Bayesian Computation in hydrologic modeling: Equifinality of formal and informal approaches, *Hydrol. Earth Syst. Sci. Discuss.*, *10*, 4739–4797, doi:10.5194/hessd-10-4739-2013.

Schoups, G., and J. A. Vrugt (2010), A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic and non-Gaussian errors, *Water Resour. Res.*, *46*, W10531, doi:10.1029/2009WR008933.

Searcy, J. K. (1959), Flow-duration curves, in *Manual of Hydrology: Part 2, Low-Flow Techniques,* U.S. Geol. Survey Water-Supply Paper 1542–A, USGS, Washington, D. C.

Shamir, E., B. Imam, E. Morin, H. V. Gupta, and S. Sorooshian (2005), The role of hydrograph indices in parameter estimation of rainfall-runoff models, *Hydrol. Processes*, *19*, 2187–2207.

Sisson, S. A., Y. Fan, and M. M. Tanaka (2007), Sequential Monte Carlo without likelihoods, *Proc. Natl. Acad. Sci. U. S. A.*, *104*(6), 1760–1765.

Smith, T., A. Sharma, L. Marshall, R. Mehrotra, and S. Sisson (2010), Development of a formal likelihood function for improved Bayesian inference of ephemeral catchments, *Water Resour. Res.*, *46*, W12551, doi:10.1029/2010WR009514.

Thyer, M., B. Renard, D. Kavetski, G. Kuczera, S. W. Franks, and S. Srikanthan (2009), Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis, *Water Resour. Res.*, *45*, W00B14, doi:10.1029/2008WR006825.

Turner, B. M., and T. van Zandt (2012), A tutorial on approximate Bayesian computation, *J. Math. Psychol.*, *56*, 69–85.

van Genuchten, M. T. (1980), A closed-form equation for predicting the hydraulic conductivity of unsaturated soils, *Soil Sci. Soc. Am. J.*, *44*(5), 892–898.

Vrugt, J. A., and B. A. Robinson (2007), Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging, *Water Resour. Res.*, *43*, W01411, doi:10.1029/2005WR004838.

Vrugt, J. A., C. G. H. Diks, W. Bouten, H. V. Gupta, and J. M. Verstraten (2005), Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation, *Water Resour. Res.*, *41*, W01017, doi:10.1029/2004WR003059.

Vrugt, J. A., C. J. F. ter Braak, M. P. Clark, J. M. Hyman, and B. A. Robinson (2008a), Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation, *Water Resour. Res.*, *44*, W00B09, doi:10.1029/2007WR006720.

Vrugt, J. A., C. J. F. ter Braak, H. V. Gupta, and B. A. Robinson (2008b), Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling?, *Stochastic Environ. Res. Risk Assess.*, *23*(7), 1011–1026, doi:10.1007/s00477-008-0274-y.

Vrugt, J. A., C. J. F. ter Braak, C. G. H. Diks, D. Higdon, B. A. Robinson, and J. M. Hyman (2009a), Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling, *Int. J. Nonlinear Sci. Numer. Simul.*, *10*(3), 273–290.

Vrugt, J. A., C. J. F. ter Braak, C. G. H. Diks, and G. Schoups (2012), Hydrologic data assimilation using particle Markov chain Monte Carlo simulation: Theory, concepts and applications, *Adv. Water Resour.*, *51*, 457–478, doi:10.1016/j.advwatres.2012.04.002.

Westerberg, I. K., J. L. Guerrero, P. M. Younger, K. J. Beven, J. Seibert, S. Halldin, J. E. Freer, and C. Y. Xu (2011), Calibration of hydrological models using flow-duration curves, *Hydrol. Earth Syst. Sci.*, *15*, 2205–2227, doi:10.5194/hess-15–2205-2011.

Yilmaz, K. K., H. V. Gupta, and T. Wagener (2008), A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, *Water Resour. Res.*, *44*, W09417, doi:10.1029/2007WR006716.

Young, P. C. (2012), Hypothetico-inductive data-based mechanistic modeling of hydrological systems, *Water Resour. Res.*, *49*,doi:10.1002/wrcr.20068, in press.