UC Irvine
UC Irvine Electronic Theses and Dissertations

Title
Towards Improved Model Evaluation: Diagnostic, Likelihood-free, Bayesian Inference

Permalink
https://escholarship.org/uc/item/067809rq

Author
Sadegh, Mojtaba

Publication Date
2015

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, availalbe at https://creativecommons.org/licenses/by/4.0/

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA,
IRVINE

Towards Improved Model Evaluation: Diagnostic, Likelihood-free, Bayesian Inference

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Civil and Environmental Engineering

by

Mojtaba Sadegh

Dissertation Committee:
Professor Jasper A. Vrugt, Chair
Professor Amir Aghakouchak
Professor Soroosh Sorooshian

2015

Chapter 2 © 2013 American Geophysical Union (AGU)
Chapter 3 © 2013 European Geophysical Union (EGU)
Chapter 4 © 2014 American Geophysical Union (AGU)
Chapter 5 © 2015 American Geophysical Union (AGU)
Chapter 6 © 2015 American Geophysical Union (AGU)
Chapter 7 © 2015 Oxford University Press (OUP)
All other materials © 2015 Mojtaba Sadegh

DEDICATION

To
my greatest, passionate, loving mom,
kind, wonderful, warm dad,
and beautiful, caring, kind-hearted sister.

TABLE OF CONTENTS

Page

LIST OF FIGURES **vi**

LIST OF TABLES **xviii**

LIST OF ALGORITHMS **xx**

ACKNOWLEDGMENTS **xxi**

CURRICULUM VITAE **xxiii**

ABSTRACT OF THE DISSERTATION **xxvii**

1 Introduction **1**
- 1.1 Traditional Model-Data Fusion . . . 1
- 1.2 Likelihood-Free Inference: Approximate Bayesian Computation . . . 7
- 1.3 Example . . . 11
- 1.4 Summary statistics versus residual-based likelihood functions . . . 13
- 1.5 Scope . . . 13

2 Towards Diagnostic Model Calibration and Evaluation: Approximate Bayesian Computation **18**
- 2.1 Introduction and scope . . . 19
- 2.2 Approximate Bayesian Computation . . . 24
- 2.3 Case Study I: Toy Problem . . . 26
- 2.4 Case Study II: Conceptual Hydrologic Modeling . . . 30
- 2.5 Discussion . . . 36
- 2.6 Summary and Conclusions . . . 37
- 2.A Appendix A . . . 39

3 Approximate Bayesian Computation: Bridging the Gap Between Formal and Informal Statistical Approaches **40**
- 3.1 Introduction . . . 41
- 3.2 Bayesian Inference . . . 48
- 3.3 Approximate Bayesian Computation . . . 51
- 3.4 Statistical Equivalence of ABC and GLUE . . . 57
- 3.5 Case Studies: Hydrologic Modeling . . . 60

3.6 Discussion and Conclusions . . . 74
3.A Appendix A . . . 79
3.B Appendix B . . . 81

4 Approximate Bayesian Computation Using Markov Chain Monte Carlo Simulation: DREAM$_{(ABC)}$ 82
4.1 Introduction and Scope . . . 83
4.2 Approximate Bayesian Computation . . . 88
4.2.1 Rejection Algorithm . . . 89
4.2.2 Population Monte Carlo Simulation . . . 90
4.3 Markov Chain Monte Carlo Simulation . . . 94
4.3.1 Continuous fitness kernel . . . 95
4.3.2 Pseudo-code of DREAM$_{(ABC)}$ . . . 99
4.3.3 Randomized subspace sampling . . . 102
4.4 Numerical experiments . . . 103
4.4.1 Synthetic benchmark experiments: Gaussian mixture model . . . 104
4.4.2 Synthetic benchmark experiments: 20-dimensional bivariate distribution 105
4.4.3 Hydrologic modeling: Sacramento Soil Moisture Accounting Model . 108
4.5 Summary and Conclusions . . . 121

5 Diagnosing Nonstationary Watershed Behavior from Process-Based Model Evaluation using Approximate Bayesian Computation 123
5.1 Introduction . . . 124
5.2 Hydrologic Data and Watersheds . . . 130
5.3 Approximate Bayesian Computation . . . 131
5.3.1 Posterior sampling: DREAM$_{(ABC)}$ . . . 135
5.3.2 Choice of Summary Metrics . . . 137
5.3.3 Choice of ϵ, the Cutoff Threshold . . . 144
5.4 Results and discussion . . . 145
5.5 Summary and conclusions . . . 154

6 The Soil Water Characteristic as New Class of Closed-form Parametric Expressions for the Flow Duration Curve 157
6.1 Introduction . . . 158
6.2 Materials and methods . . . 164
6.2.1 Experimental data: MOPEX data set . . . 164
6.2.2 Probabilistic and mathematical functions . . . 166
6.2.3 Proposed formulations . . . 168
6.2.4 FDC model parameter estimation . . . 172
6.3 Results and discussion . . . 174
6.4 Conclusions . . . 191

7 Summary Statistics From Training Images as Prior Information in Probabilistic Inversion 193
7.1 Introduction . . . 194

7.2 Methodology 197
7.2.1 Sparse model parameterization 197
7.2.2 Markov chain Monte Carlo sampling with DREAM(ZS) 199
7.2.3 Summary statistics from training image realizations 203
7.2.4 Prior distribution based on summary statistics 206
7.2.5 Forward modeling 209
7.3 Results 210
7.3.1 Channels case 211
7.3.2 Fluvial deposits case 218
7.4 Discussion 222
7.5 Conclusions 225

8 Conclusions **226**

Bibliography **231**

LIST OF FIGURES

Page

1.1 Explicit recognition of the role of 1: parameter, 2: forcing data, 3: initial state, 4: model structural, 5: output, 6: state, and 7: calibration data uncertainty. The pitchfork symbol illuminates the difficulty which formulation of the likelihood function (and prior distribution / parameterization of individual error sources) to use to summarize the error residuals. Explicit treatment of individual error sources is required to increase the prospects of explaining the reasons for model inadequacy and learning from the experimental data [*Sadegh and Vrugt*, 2013]. 3

1.2 Conceptual overview of approximate Bayesian computation (ABC) for a hypothetical one-dimensional parameter estimation problem. First, N samples are drawn from a user-defined prior distribution, $\theta^* \sim p(\theta)$. Then, this ensemble is evaluated with the model and creates N model simulations. If the distance between the observed and simulated data, $\rho(\tilde{\mathbf{Y}}, \mathbf{Y}(\theta^*))$ is smaller than or equal to some nominal value, ϵ then θ^* is retained, otherwise the simulation is discarded. The accepted samples are then used to approximate the posterior parameter distribution, $p(\theta|\tilde{\mathbf{Y}})$. Note that for sufficiently complex models and large data sets the probability of happening upon a simulation run that yields precisely the same simulated values as the observations will be very small, often unacceptably so. Therefore, $\rho(\tilde{\mathbf{Y}}, \mathbf{Y}(\boldsymbol{\theta}^*))$ is typically defined as a distance between summary statistics of the simulated, $S(\mathbf{Y}(\boldsymbol{\theta}^*))$ and observed, $S(\tilde{\mathbf{Y}})$ data, respectively. Modified after *Sunnåker et al.* [2013]. 9

1.3 Synthetic data (solid circles) of the toy example and the original linear regression function, $Y \leftarrow \frac{1}{2}t+5, t \in \{0, 10\}$ (solid black line) that was used to create the observations. The ABC framework now compares summary statistics of the measurement and simulated data to estimate the slope and intercept of the regression function. 11

1.4 Marginal posterior distribution of the slope, θ_1 and intercept, θ_2 derived from least squares fitting (top panel) and ABC inference using one (mean of data) (middle panel) and two (mean and standard deviation of data (bottom panel) summary diagnostic(s). 12

2.1 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 $\bigotimes$ 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. 22
2.2 Synthetic data (solid circles) of the toy example and the original linear regression function, $Y \leftarrow \frac{1}{2}t+5, t \in \{0, 10\}$ (solid black line) that was used to create the observations. The ABC framework now compares summary statistics of the measurement and simulated data to estimate the slope and intercept of the regression function. 27
2.3 Marginal posterior distribution of the slope, θ_1 and intercept, θ_2 derived from least squares fitting (top panel) and ABC inference using one (mean of data) (middle panel) and two (mean and standard deviation of data (bottom panel) summary diagnostic(s). 27
2.4 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 $\epsilon = 0.1$ generates stable results. 29
2.5 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. 32
2.6 Observed (red dots) and model fitted (solid black line) flow duration curve (FDC) of the French Broad River basin at Asheville, North Carolina, USA. The two-parameter model closely matches the observed FDC. The optimized parameters, S_3 and S_4 of the FDC model are used as hydrologic signatures in our ABC analysis. 34

2.7 ABC inference for the French Broad River basin using summary metrics $S_1 \rightarrow S_4$ that represent the observed annual runoff coefficient, annual baseflow 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 non-driven part of the hydrograph. 35

3.1 Explicit recognition of the role of 1: parameter, 2: forcing data, 3: initial state, 4: model structural, 5: output, 6: state, and 7: calibration data uncertainty. The pitchfork symbol illuminates the difficulty which formulation of the likelihood function (and prior distribution / parameterization of individual error sources) to use to summarize the error residuals. Explicit treatment of individual error sources is required to increase the prospects of explaining the reasons for model inadequacy and learning from the experimental data. . 42

3.2 Synthetic discharge time series (blue line) simulated with the Nash Cascade model, and the corrupted observations (red points) used in the GLUE and ABC analysis. 54

3.3 Posterior distribution of the recession parameter derived from (A) ABC using the mean of the data as summary metric, (B) DREAM using a heteroscedastic measurement error, and (C) ABC using the mean and standard deviation of the data as summary statistics. The true parameter value is indicated with a red cube on top of each plot. 56

3.4 95% streamflow uncertainty ranges (dark region) derived from GLUE (top panel) and ABC (bottom panel). The red points mark the actual discharge observations. The prediction uncertainty ranges derived with both methods appear virtually identical and nicely capture the desired percentage of streamflow observations. 60

3.5 Posterior distribution of five randomly chosen SAC-SMA model parameters derived from (A) GLUE, and (B) ABC using historical streamflow data from the French Broad river basin. 65

3.6 SAC-SMA derived 95% streamflow simulation uncertainty ranges (grey region) of the calibration period of the French Broad river basin using (A) GLUE and (B) ABC. The observed discharge data are indicated with the solid red dots. We limit our display to the first 365-days of the calibration data set to simplify graphical interpretation. The simulation uncertainty ranges appear very similar and nicely cover the observed discharge data. 65

3.7 Bivariate scatter plots of the behavioral (posterior) samples of three different (randomly selected) parameter pairs using GLUE (top panel) and ABC (bottom panel): (A) LZSK-LZPK, (B) PFREE-ADIMP, and (C) ZPERC-LZPK. The scatter plots derived with both methods are in close agreement but demonstrate an important difference in sampling density. The computational budget of GLUE was limited to approximately 2 days, and within this time frame the Latin hypercube sampling method located only 300 behavioral solutions. 67
3.8 SAC-SMA derived 95% streamflow simulation uncertainty ranges (grey region) for a three-year portion of the evaluation period of the French Broad river basin using the (A) GLUE and (B) ABC derived posterior parameter distribution. The observed discharge data are indicated with the solid red dots. 68
3.9 Posterior distribution of five randomly selected hmodel parameters derived from the French Broad calibration data set, (A) $I_{\max}$, (B) $Q_{\max}$, (C) α_E, (D) α_F, and (E) K_F. 70
3.10 Bivariate scatter plots of the behavioral (posterior) samples of three different (randomly selected) hmodel parameter pairs using GLUE (top panel) and ABC (bottom panel): (A) $\alpha_F - Q_{\max}$, (B) $K_F - K_S$, and (C) $S_{\max} - K_F$. The scatter plots derived with both methods are in close agreement but demonstrate an important difference in sampling density. The computational budget of GLUE was limited to about 3 days, and this has resulted in 100 behavioral solutions. 71
3.11 Hmodel derived 95% streamflow simulation uncertainty ranges (grey region) of the calibration period of the French Broad river basin using (A) GLUE and (B) ABC. The observed discharge data are indicated with the solid red dots. We limit our display to the first 365-days of the calibration data set to facilitate graphical interpretation. The uncertainty ranges appear very similar and nicely cover the observed discharge data. 72
3.12 Hmodel derived 95% streamflow simulation uncertainty ranges (grey region) for a three-year portion of the evaluation period of the French Broad river basin using the (A) GLUE and (B) ABC derived posterior parameter distribution. The observed discharge data are indicated with the solid red dots. . 73
3.13 Posterior distribution of five randomly chosen SAC-SMA model parameters derived from GLUE (top panel), and ABC (bottom panel) using historical streamflow data from the Leaf River watershed in Mississippi. 74
3.14 SAC-SMA derived 95% streamflow simulation uncertainty ranges (grey region) for a three-year portion of the evaluation period of the Leaf River watershed using the (A) GLUE and (B) ABC derived posterior parameter distribution. The observed discharge data are indicated with the solid red dots. 75

3.15 Illustration of the effect of measurement data uncertainty on the ABC-derived posterior parameter and streamflow uncertainty for a representative 250-day portion of the French Broad calibration data set: (A) 95% simulation uncertainty assuming a deterministic, $\mathbf{Y} \leftarrow \mathcal{H}(\boldsymbol{\theta}|\cdot)$, and (D) stochastic model operator, $\mathbf{Y} \leftarrow \mathcal{H}(\boldsymbol{\theta}|\cdot) + N(0, \hat{\sigma}_{\tilde{\mathbf{Y}}})$ with discharge measurement error derived from the nonparametric estimator in equation (3.11). The observed discharge data are indicated with the solid red dots. The histograms at the right-hand side of each individual panel plot the corresponding marginal posterior distributions of the SAC-SMA model parameters ADIMP and PCTIM. 79

4.1 Conceptual overview of approximate Bayesian computation (ABC) for a hypothetical one-dimensional parameter estimation problem. First, N samples are drawn from a user-defined prior distribution, $\theta^* \sim p(\theta)$. Then, this ensemble is evaluated with the model and creates N model simulations. If the distance between the observed and simulated data, $\rho(\tilde{\mathbf{Y}}, \mathbf{Y}(\theta^*))$ is smaller than or equal to some nominal value, ϵ then θ^* is retained, otherwise the simulation is discarded. The accepted samples are then used to approximate the posterior parameter distribution, $p(\theta|\tilde{\mathbf{Y}})$. Note that for sufficiently complex models and large data sets the probability of happening upon a simulation run that yields precisely the same simulated values as the observations will be very small, often unacceptably so. Therefore, $\rho(\tilde{\mathbf{Y}}, \mathbf{Y}(\boldsymbol{\theta}^*))$ is typically defined as a distance between summary statistics of the simulated, $S(\mathbf{Y}(\boldsymbol{\theta}^*))$ and observed, $S(\tilde{\mathbf{Y}})$ data, respectively. Modified after *Sunnåker et al.* [2013]. 87
4.2 One-dimensional mixture distribution (solid black line) and histogram of the posterior samples derived from (A) Rejection sampling (ABC-REJ) , (B) Population Monte Carlo sampling (ABC-PMC), and (C) MCMC simulation with DREAM$_{(\text{ABC})}$. 93
4.3 Two-dimensional scatter plots of the posterior samples, "+" generated with (A) Population Monte Carlo sampling, and (B) MCMC simulation with DREAM$_{(\text{ABC})}$. The true values of μ_1 and μ_2 are indicated with a cross, "×". The ABC-PMC samples are very dispersed and significantly over estimate the actual width of the target distribution. 106
4.4 Trace plots of the $\hat{R}$-statistic of the sampled Markov chains with DREAM$_{(\text{ABC})}$ for the 20-dimensional bivariate normal distribution. Each parameter is coded with a different color. The dashed line denotes the default threshold used to diagnose convergence to a limiting distribution. 108

4.5 Marginal distributions of PCTIM, ADIMP, LZFSM, LZFPM and LZPK derived from the posterior samples created with ABC-PMC (upper panel) and DREAM$_{(ABC)}$ (middle panel). The histograms of the five SAC-SMA parameters occupy almost their entire prior distribution, which suggests that they are not particularly well identifiable by calibration against the observed annual baseflow index, annual runoff coefficient, and flow duration curve, respectively. The results of both ABC sampling methods are in good (visual) agreement, which inspires confidence in the ability of DREAM$_{(ABC)}$ to correctly sample the underlying target distribution. The bottom panel displays histograms of the SAC-SMA parameters derived using a classical residual-based (Gaussian) likelihood function. The SAC-SMA parameters are much better resolved, but these results cannot be justified given (amongst others) a lack of treatment of rainfall data errors. 112

4.6 95% posterior simulation uncertainty ranges (gray region) of the SAC-SMA model for a selected portion of the evaluation data set of the French Broad watershed. The top two panels displays the results of diagnostic model evaluation using (A) ABC-PMC, and (B) DREAM$_{(ABC)}$, whereas the bottom panel depicts the results of DREAM with a classical residual-based Gaussian likelihood function. The observed discharge values are indicated with the red dots. The SAC-SMA simulation intervals derived from ABC-PMC and DREAM$_{(ABC)}$ are very similar and encapsulate a large part of the discharge observations. The DREAM$_{(RBGL)}$ uncertainty ranges, on the other hand, exhibit a much lower coverage, but closer track the observed discharge data. 114

4.7 Evolution of the $\hat{R}$-statistic for the parameters of the SAC-SMA model using DREAM$_{(ABC)}$ and discharge data from the French Broad watershed. Each of the parameters is coded with a different color. The dashed line denotes the default threshold used to diagnose convergence to a limiting distribution. . . 118

4.8 Histograms of the SAC-SMA derived summary statistics, S_1 (runoff), S_2 (baseflow), S_3 and S_4 (flow duration curve) of the posterior samples from ABC-PMC (top panel) and DREAM$_{(ABC)}$ (middle panel). The bottom panel displays the results of DREAM with a residual-based Gaussian likelihood function. The observed values of the summary metrics are separately indicated in each plot using the "×" symbol. While $S_2 \rightarrow S_4$ center around their observed value (×) for the ABC analysis, the marginal posterior distribution of S_1 is skewed to the right and does not encapsulate its measured value. This demonstrates that model is unable to simultaneously satisfy all the four different summary metrics used herein. This points to a structural deficiency in the SAC-SMA model structure, which will be investigated in more detail in subsequent work. 119

4.9 Two-dimensional scatter plots of the posterior samples, "+" generated with DREAM$_{(ABC)}$ for six different pairs of parameters of the SAC-SMA model. We restrict our attention to the (A) PCTIM - LZPK, (B) LZSK - LZPK, (C) PCTIM - LZFPM, (D) PCTIM - LZSK, (E) LZPK - PFREE, and (F) ADIMP - LZPK space, respectively. The bivariate posterior samples occupy a well-defined hypercube interior to the prior distribution. 120

5.1 Synthetic time series of data generated by a stochastic stationary process to demonstrate the importance of considering time scales in nonstationary behavior analysis. The top panel provides a snapshot of 200 years of data from the original time series of 2,000 years (bottom panel). Artificial data is represented with a blue line, whereas the red line represents the mean system response. The top panel demonstrates a sudden shift in system response after exactly 100 years of observation. This abrupt change of the system response would seem a textbook example of nonstationary, yet when the behavior of the system is viewed over a much longer, 2,000 year data period, the mean and standard deviation of $\tilde{\mathbf{Y}}$ appear stable, and indeed stationarity is observed. One should therefore be particularly careful proclaiming nonstationarity on the basis of graphical and/or simple statistical analysis, and a relatively small data set. Note, this example is adapted from *Koutsoyiannis* [2011]. 128

5.2 Cumulative streamflow (a) and precipitation (b) data record of the Blackberry creek in Illinois, USA for the years 2002-2012. The slope of the cumulative streamflow curve increases noticeably between the years 2007 and 2008. In the hydrologic community this behavior is classified as nonstarionary due to growing urbanization (`http://non-stationarities.irstea.fr/`). Yet, the bottom graph demonstrates noticeably larger precipitation amounts from 2007 onwards which coincides perfectly with the increase in streamflow. With the help of diagnostic model evaluation and temporal analysis of the hydrologic signatures, we classify this watershed as stationary. A single model structure and fixed parameterization is sufficient to model the rainfall-runoff response. 129

5.3 Schematic overview of likelihood-free (ABC) inference for a model involving only one parameter using two different summary metrics, S_1 and S_2 of the measured data, $\tilde{\mathbf{Y}}$. The original watershed data, $\tilde{\mathbf{Y}}$ is plotted in step 1, and is used to calculate the observed summary metrics (step 2). After the user has defined the value of ϵ, step 3 proceeds by drawing N samples from the prior parameter distribution, $p(\theta)$. In step 4, each of these N samples is evaluated by the model, and used to calculate the simulated values of the summary metrics. Step 5 then proceeds with a comparison of the observed and simulated summary metrics. Model realizations whose simulated summary metrics fall within the purple box, and thus satisfy $\rho(S(\mathbf{Y}(\boldsymbol{\theta}_i^*)), S(\tilde{\mathbf{Y}})) \leq \epsilon$ are used in step 6 to approximate the posterior parameter distribution. . . . 135

5.4 Histograms of the residuals between the observed and simulated summary metrics. The observed summary metrics are derived from ten different SAC-SMA discharge simulations each with a different parameterization. The simulated metrics are derived for the same SAC-SMA parameterizations but using thousand error corrupted rainfall records. The upper panel (a-e) plots the results for the least sensitive summary metrics, whereas the bottom panel (f-j) considers metrics that are much more sensitive to rainfall data errors. 143

5.5 Marginal posterior parameter distributions of the SAC-SMA model parameters PCTIM, ADIMP, LZTWM, LZPK, and LZSK derived from DREAM$_{(ABC)}$ using historical data from the Oostanaula (top), French Broad (middle) and Leaf river (bottom) watersheds, respectively. Some parameters are very well defined by calibration against the 14 summary metrics, whereas others exhibit considerable uncertainty. 147
5.6 Trace plots of the sampled distance values, $\rho(\cdot)$, in each of the chains for the Oostanaula (a), French Broad (b) and Leaf (c) river watersheds. The dashed red line depicts the threshold of $\epsilon = 0.05$ used to differentiate between behavioral and nonbehavioral samples. 148
5.7 Histograms of the marginal posterior distributions of the runoff coefficient (S_1), baseflow index (S_2), FDC (S_3 and S_4) and rising limb density for the 7th year of the calibration period for the Oostanaula (up), French Broad (middle) and Leaf (bottom) river watersheds. The symbol "x" denotes the measured values of the summary metrics. These plots confirm that each of the posterior summary metrics lie within $\epsilon = 0.05$ (vertical dashed red lines) of their observed counterparts. 149
5.8 95% posterior SAC-SMA simulation uncertainty ranges for the Oostanaula (a), French Broad (b) and Leaf (c) river watersheds for a representative 365 days portion of the calibration data period. The simulation uncertainty ranges (grey region) envelop between 37-61% of the streamflow observations (red dots). 150
5.9 Trace plots of the sampled distance values, $\rho(\cdot)$, in each of the three Markov chains simulated with the DREAM$_{(ABC)}$ algorithm for the Wights catchment (a), and the synthetic watersheds "abrupt" (b) and "gradual" (c). The dashed red line depicts the threshold of $\epsilon = 0.05$ used to distinguish between behavioral and nonbehavioral samples. The inability to find samples within epsilon of the measured summary metrics is interpreted as an incompatibility between the stationary assumption of the model structure and/or parameterization used, and the nonstationary behavior observed in the underlying data. . . . 151
5.10 Histograms of the marginal posterior distributions of the runoff coefficient (S_1), baseflow index (S_2), FDC (S_3 and S_4) and rising limb density for the 7th year of the calibration period of the Wights catchment (top), and synthetic watersheds "gradual" (middle) and "abrupt" (bottom). The symbol "x" marks the measured values of the summary metrics. The "posterior" simulations of the SAC-SMA model cannot be considered behavioral as the summary metrics deviate too much from their observed counterparts. A single parameterization of the SAC-SMA model is unable to accurately describe each of the 14 observed summary metrics for each of the three basins. These results provide evidence that the watersheds behavior cannot be described adequately with a single model parameterization and hence has experienced a change during the calibration data period. 153

6.1 Flow duration curves for eight watersheds from the MOPEX data set, which show contrasting hydrologic behaviors. This figure plots normalized streamflows (mm/day) versus exceedance probability (-). We normalized the streamflow values to plot various shapes of the FDCs in one figure. 165
6.2 Water retention curves for five different soil types introduced in *van Genuchten* [1980]. This figure plots soil water pressure head (cm) versus soil water content (cm^3/cm^3). Note that the soil water pressure head values are originally negative, but we use positive pressure heads for plotting since the absolute values are used in the WRFs. 169
6.3 Histogram of the root mean square errors (RMSEs) between the simulated and observed FDCs for the 408 studied watersheds. The x-axis in each plot represents the RMSE values (dimensionless) and the y-axis shows the density of the number of occurrences for each RMSE value. Left, middle and right columns show the results of 2-, 3-, and 5-parameter model formulations, respectively. The results of the VG (plots: A, E, I), Kosugi (plots: B, F, J), Gumbel (plot: C), Quimpo (plot: D), GEV (plot: G), and Franchini and Suppo (plot: H) are presented with black, blue, cyan, dark green, light green, and gray bins, respectively. The RMSE shows the deviation between the simulated and empirical exceedance probabilities, and is hence dimensionless. 177
6.4 Partial parameter sensitivities of the VG (top panel) and Kosugi (bottom panel) models of the FDC (see Table 6.1) for a range of different streamflow values. Color coding is used to discern between the parameters a_{x} and $a_{1,\mathrm{x}}$ (red), b_{x} and $b_{1,\mathrm{x}}$ (blue), c_{x} and $a_{2,\mathrm{x}}$ (green), $b_{2,\mathrm{x}}$ (cyan), and $w_{1,\mathrm{x}}$ (black) of the different model formulations. The subscript X stands for VG and K. To help with visual interpretation, the gray dashed line depicts, for all flow levels, a zero sensitivity. 178
6.5 Comparison of the observed (red dots) and fitted (different lines) FDCs of the Green river watershed near Genesco, IL, USA. The top (A-C) panel uses a linear scale of the streamflow values, whereas the bottom (D-F) panel uses a logarithmic scale. The different columns represent different model complexities. Color coding is used to differentiate between the FDC models of Table 6.1: VG (black), Kosugi (blue), Gumbel (cyan), Quimpo (dark green), GEV (light green) and Franchini and Suppo (gray). 181
6.6 Comparison of the observed (red dots) and fitted (different lines) FDCs of the Kankakee river at Davis, IN, USA. The top (A-C) panel uses a linear scale of the streamflow values, whereas the bottom (D-F) panel uses a logarithmic scale. The different columns display the results of the 2 (left), 3 (middle) and 5 (right) parameter formulations.Color coding is used to differentiate between the FDC models of Table 6.1: VG (black), Kosugi (blue), Gumbel (cyan), Quimpo (dark green), GEV (light green) and Franchini and Suppo (gray). . 182

6.7 Comparison of the observed (red dots) and fitted (different lines) FDCs of the Little river near Horatio, AR, USA. A linear scale of the streamflow values is used in the top panel, whereas the bottom panel uses a logarithmic y-scale. The different columns display the results of the 2 (left), 3 (middle) and 5 (right) parameter formulations. The VG model is portrayed in black, and the Kosugi model in blue. The other FDC models proposed in the literature are displayed in cyan (Gumbel), dark green (Quimpo), light green (GEV) and gray (Franchini and Suppo). 183
6.8 Comparison of the simulated FDCs of the 2-parameter Kosugi model (blue lines) against their observed counterparts (red dots) of the MOPEX data set. The different graphs correspond to the (A) SB Potomac, (B) Tygart Valley, (C) NF Holston, (D) Little, (E) Green, (F) Kankakee, (G) EF San Gabriel, (H) White, (I) Pemigewasset, (J) Little Pee Dee, (K) Licking, and (L) Genesee river basins, respectively, which are found in the states of West Virginia (A,B), Virginia, Arkansas, Illinois, Indiana, California, Washington, New Hampshire, South Carolina, Kentucky, and New York. 186
6.9 Comparison of the simulated FDCs of the 5-parameter Kosugi model (blue lines) against their empirical counterparts (red dots) of the MOPEX data set. The different graphs pertain to the (A) SB Potomac, (B) Tygart Valley, (C) NF Holston, (D) Little, (E) Green, (F) Kankakee, (G) EF San Gabriel, (H) White, (I) Pemigewasset, (J) Little Pee Dee, (K) Licking, and (L) Genesee river basins respectively which are located in the states of West Virginia (A,B), Virginia, Arkansas, Illinois, Indiana, California, Washington, New Hampshire, South Carolina, Kentucky, and New York. 187
6.10 95% simulation uncertainty ranges of the 5-parameter Kosugi model for the twelve different watersheds of Figure 6.8. The observed data are indicated with the red dots, whereas the 95% parameter and total uncertainty are displayed using the dark and light gray regions, respectively. The fit to the empirical FDCs is excellent and the parameter and model uncertainty therefore negligibly small. 188
6.11 Scatter plot of first parameter of the original Kosugi FDC model, a_{K}, vs median streamflow, y_{Mdn}, for 408 watersheds of the MOPEX data set. a_{K} and y_{Mdn} are correlated with a linear correlation coefficient $r = 0.9974$ and a line $y_{\mathrm{Mdn},i} = 1.0061 a_{\mathrm{K},i} - 0.0061$ describes the relationship between them. 189

7.1 Schematic workflow of the presented inversion strategy. Realizations of a training image (TI) generated by multiple-point statistics (MPS) simulations are subject to discrete cosine transformation (DCT). Dominant DCT coefficients are derived and the corresponding basis functions are used to parameterize the porosity fields. From compressed TI realizations, summary statistics distributions are extracted and translated into a prior probability density function. The model space is sampled with the DREAM$_{(\mathrm{ZS})}$ algorithm to obtain posterior model realizations. 204

7.2 Formulation of the summary metrics prior and the indicator function. The distribution of roughness values observed in a set of 1,000 TI realizations is shown in gray and the estimated prior following this distribution is shown by the black solid line. The blue bars depict the distribution of roughness values in an initial set of random models. To ensure that the posterior models have roughness values within the summary metric prior range, an indicator function (dashed black line) is applied. The indicator function is 1 where the summary metrics prior is non-zero and it decreases to 0 towards the possible extreme values that might be sampled. Only realizations for which the indicator function is 1 are considered as possible realizations of the posterior. 207
7.3 Training images for the channels case (a), the fluvial deposits case (c) and multiple-point statistics realizations thereof (b and d). Note that the full ensemble of TI realizations contains 1,000 realizations for both cases. 211
7.4 Channels case: a) reference porosity field with GPR transmitter (left side) and receiver (right side) antenna locations indicated by white dots, b) compressed image of the reference porosity field obtained by retaining the $n = 100$ highest amplitude DCT coefficients of the reference field. c) Ensemble mean of the summary metrics prior (variability and porosity percentiles) using the TI-based parameterization obtained by running an MCMC simulation with the likelihood set to unity. d) and e) Mean of the inferred porosity posterior without consideration of a summary metrics prior without (d) and with (e) TI-based parameterization. f)-i) Posterior mean of models obtained with TI-based parameterization and considering a summary metrics prior using f) roughness metric, g) variability metric, h) porosity percentile metrics and i) both variability and percentile metrics. 212
7.5 a) Random samples from the ensemble of summary metrics prior (variability and porosity metrics) models using the TI-based parameterization for the channels case. Corresponding random samples from the ensemble of posterior models for the channels case: b) no TI-based parameterization, no summary metric prior, c) TI-based parameterization, no summary metric prior, d)-g) TI-based parameterization, summary metric prior using d) roughness metric, e) variability metric, f) porosity percentile metrics, g) both variability and percentile metrics. 214
7.6 Posterior distributions of the petrophysical parameters ϵ_s and m, and of the relative data error level for the channels case: a) no TI-based parameterization, no summary metric prior, b) TI-based parameterization, no summary metric prior, c)-f) TI-based parameterization, summary metric prior using c) roughness metric, d) variability metric, e) percentile metrics, f) both variability and percentile metrics. Red crosses depict the reference values. 215
7.7 Maps of the posterior log-probability for all possible combinations of the petrophysical parameters m and ϵ_s for the channels case. The DCT coefficients are fixed at their maximum a-posteriori density (MAP) values. Indicated are the parameter values of the reference (red crosses), the values for which the log-probability is maximized (black crosses) and the estimate found by DREAM(ZS) (blue asterisks). Enumerations correspond to those of Figure 7.6.218

7.8 Summary metric prior distributions (black lines, see Figure 7.2 for explanation) and the posterior distributions of the summary metrics (blue bars) for the channels case: a) roughness metric, b) variability metric, c)-e) 10, 50, 90% percentile metrics. Red crosses display the summary metrics of the reference porosity field. 218
7.9 Fluvial deposits case: a) reference porosity field with GPR antenna locations indicated by white dots, b) compressed image of the reference porosity field obtained by retaining the $n = 150$ highest amplitude DCT coefficients of the reference field. c) Ensemble mean of the summary statistics prior (roughness and porosity percentiles) using the TI-based parameterization obtained by running an MCMC simulation with the likelihood always equal to 1. d) and e) Mean of the inferred porosity posterior without consideration of a summary metric prior without (d) and with (e) TI-based parameterization. f) Posterior mean of models with TI-based parameterization considering a summary metric prior based on roughness and porosity percentiles. 219
7.10 a) Random samples from the ensemble of summary statistics prior (roughness and porosity percentiles) using the TI-based parameterization for the fluvial deposits case. Corresponding random samples from the ensemble of posterior models: b) no TI-based parameterization, no summary metric prior, c) TI-based parameterization, no summary metric prior, d) TI-based parameterization, summary metric prior using roughness and porosity percentile metrics. 220
7.11 Posterior distributions of the petrophysical parameters ϵ_s and m, and of the relative data error level for the fluvial deposits case: a) no TI-based parameterization, no summary metric prior, b) TI-based parameterization, no summary metric prior, c) TI-based parameterization, summary metric prior using roughness and percentile metrics. Red crosses depict the reference values. . . 221
7.12 Maps of the posterior log-probability for all possible combinations of the petrophysical parameters m and ϵ_s for the fluvial deposits case. The DCT coefficients are fixed at their maximum a-posteriori density (MAP) values. Indicated are the parameter values of the reference (red crosses), the values for which the log-probability is maximized (black crosses) and the estimate found by DREAM$_{(ZS)}$ (blue asterisks). Enumerations correspond to those of Figure 7.11. 221
7.13 Prior probability density functions derived from the distribution of summary metrics in the ensemble of TI realizations (black lines, see Figure 7.2 for explanation) and the posterior distributions of the summary metrics (blue bars) for the fluvial deposits case: a) roughness metric, b)-d) 10, 50, 90% percentile metrics. Red crosses depict the summary metrics of the reference porosity field. 222

LIST OF TABLES

Page

2.1 Prior Uncertainty Ranges of Hydrologic Model Parameters. 30

3.1 Prior uncertainty ranges of hmodel parameters. 61
3.2 Description of the SAC-SMA model parameters and their (uniform) prior uncertainty ranges. 63
3.3 Computational efficiency of GLUE and ABC for the French Broad data set: Acceptance rate (AR, %) and total number of SAC-SMA and hmodel function evaluations (FE) required to sample $N = 1,000$ behavioral solutions. 66
3.4 Summary statistics of the GLUE, ABC and DREAM (formal likelihood function) derived posterior parameter distribution for the calibration and evaluation period of the French Broad river basin: Root mean square error (RMSE) of the posterior mean SAC-SMA simulation and coverage of the associated 95% streamflow simulation uncertainty ranges. 69
3.5 Summary statistics of the GLUE, ABC and DREAM (formal likelihood function) derived posterior parameter distribution for the calibration and evaluation period of the French Broad river basin: Root mean square error (RMSE) of the posterior mean hmodel simulation and coverage (%) of the associated 95% streamflow simulation uncertainty ranges. 71
3.6 Summary statistics of the GLUE, ABC and DREAM (formal likelihood function) derived posterior parameter distribution for the calibration and evaluation period of the Leaf River watershed: Root mean square error (RMSE) of the posterior mean SAC-SMA simulation and coverage (%) of the associated 95% streamflow simulation uncertainty ranges. 76

4.1 Case study I: 1-dimensional toy example. We list the (final) epsilon value, acceptance rate, AR (%) and number of function evaluations needed to sample the target distribution. 94
4.2 Case study II: 20-dimensional bivariate Gaussian distribution. We list the final epsilon value, acceptance rate, AR (%) and number of function evaluations needed to sample a stationary distribution. 106
4.3 Prior ranges of the SAC-SMA model parameters and their posterior mean values derived for the French Broad river basin data using the ABC-PMC and $\text{DREAM}_{(\text{ABC})}$ algorithms. We also summarize the results of a residual-based Gaussian likelihood function, $\text{DREAM}_{(\text{RBGL})}$. 109

4.4 Case study III: 14-dimensional SAC-SMA model calibration problem. We list the final epsilon value, acceptance rate, AR (%) and number of function evaluations needed to sample the posterior distribution. We also include the results of DREAM using a residual-based Gaussian likelihood function. . . . 111
4.5 Performance of the SAC-SMA model for the calibration and evaluation data period of the French Broad river basin. We summarize the coverage (%) and average width (mm/d) of the 95% simulation intervals (due to parameter uncertainty), and the RMSE (mm/d), bias (%) and correlation coefficient, R (-) of the posterior mean SAC-SMA simulation. 117

5.1 Description of the watersheds used herein including name, region, USGS ID and drainage area (km^2), mean annual precipitation, $\tilde{\mathbf{P}}$ (mm/day), mean annual evapotranspiration, $\mathbf{P}\tilde{\mathbf{E}}\mathbf{T}$ (mm/day), and mean annual runoff coefficient, $\tilde{\mathbf{Y}}/\tilde{\mathbf{P}}$ (-). The top ten watersheds are derived from the MOPEX data set and deemed stationary in the hydrologic literature. The other nine watersheds are classified as nonstationary. Five of these watersheds are located in Australia, two are located in the United States, and the other two constitute synthetic data sets created with the SAC-SMA model and used to validate the proposed methodology. The last column lists, for each watershed, the minimum distance between the observed and simulated summary metrics (outcome of Equation 5.5). 133
5.2 Name and description of each of the summary statistics used in our rainfall error analysis. The last column provides a reference for each metric, and can be used as guide for their numerical calculation. 140
5.3 Description of the SAC-SMA model parameters and their (uniform) prior uncertainty ranges. 142

6.1 Summary of the probablistic, mathematical and WRF-based functions of the FDC. All these functions compute the exceedance probability, G_t for a given streamflow value, y_t. The variables a, b, and c are fitting coefficients whose values need to be derived by calibration against the observed, empirical, FDC of individual watersheds. 172
6.2 Performance statistics of the quality of fit of each individual FDC model. We list the mean and standard deviation of the Root Mean Square Error (RMSE) (derived from Equation 6.12), and the number of times (in %) each respective model within a certain parametric class provides the best fit (lowest RMSE) to the observed FDCs. 175
6.3 Average linear correlation coefficients of the DREAM-derived posterior parameter samples for the 408 different watersheds of the MOPEX data set. The variable "X" that is used in subscript stands for K or VG respectively. . 184

7.1 Mean and standard deviation of the average root mean square error between posterior porosity models and the reference model for different degrees of prior information. 216

LIST OF ALGORITHMS

		Page
1	Rejection sampler	9
2	Rejection sampler (ABC-REJ)	25
3	ABC Population Monte Carlo sampler	39
4	Rejection sampler (ABC-REJ)	52
5	ABC-PMC	80
6	ABC-Rejection Sampler	89
7	ABC-Population Monte Carlo sampler	91
8	DREAM$_{(\mathrm{ABC})}$-Markov chain Monte Carlo sampler	101
9	Subspace sampling	103

ACKNOWLEDGMENTS

In September 2010, I arrived at UC Irvine as an international student from Iran with little knowledge of life in the US. Now more than four years later, I am reflecting back on a remarkable journey. I like to thank a number of people that made this journey possible, and helped me along the way. First and foremost, I would like to express my sincere gratitude to my adviser, mentor, friend, and guide, Prof. Jasper A. Vrugt. His patience, care and nurturing helped me to adapt quickly at UCI. What is more, Prof. Vrugt has provided so many insightful discussions about research, academia, and life in general. His creativity, deep knowledge and understanding of different fields of study, and close involvement with my work, enabled me to progress rapidly on different fronts and maximize my research output. His astute guidance and energetic and enthusiastic personality encouraged me to explore science more deeply, and always find the right resources to craft solutions to different research questions. I am also grateful for his generous funding support, his coding and careful editing of my papers, and for being so accessible! Beyond our daytime communications, Prof. Vrugts willingness to respond to my emails throughout the night has astonished me.

I would like to also thank my committee members, Profs. Amir Aghakouchak and Soroosh Sorooshian, for their support, flexibility and care. I remember vividly Prof. Sorooshians Hydrologic Systems class which taught me the basics of hydrologic modeling, and I also am deeply grateful for his fatherly advice, genuine care, and our scientific discussions. His life stories were a source of inspiration and joy, and his broad view on the field of hydrology has enriched my work. I would also like to thank Prof. Aghakouchak for his perceptive discussions, and for teaching me methods to analyzing extreme events. I appreciate all I learned from him (and my adviser) on developing a career in academia.

Additionally, I would like to express my gratitude to Prof. Hoshin Gupta for hosting me during a research visit to the University of Arizona. He stimulated me to think outside the box, and I appreciate his great influence on my academic and non-academic life.

I would also like to thank Prof. Reza Kerachian for guiding me toward the science, and for teaching me how to do research and publish my scientific findings. I appreciate the time and energy he spent to develop me as a researcher.

I would also like to acknowledge my friends role in my life and thank them for their support, love, passion and help. Thank you Morteza Shakeri, Reza Afshar, Mahboobe Eshkevari, Mahdieh Allahviranloo, Mahdi Navari, Nasrin Nasrollahi, Omid Esmaili, Morvarid Azizian, Nasir Emami, Phu Nguyen, Hamed Ashouri, Mahshid Fardadi, Omid Mazdiyasni, Ali Mehran, Alireza Farahmand, Vahid Haghdoust, Navid Bahramfar, and others. Moreover, thank you my office mates Aryan Safaie, Elias Massoud, Veysel Yildiz, Saeedeh Yavari, Eric Laloy, Jiri Nossent and Tobias Lochbuehler for providing such a nice and friendly environment. I would also like to honor the memory of my friend Tobias Lochbuehler who sadly passed away in a tragic accident in the Alps mountains. He was a brilliant, kind and good-natured individual.

Most importantly thank you my beautiful mom, Heshmat, my kind dad, Hossein, and my gorgeous sister, Zahra.

CURRICULUM VITAE

Mojtaba Sadegh

EDUCATION

Doctor of Philosophy in Civil and Environmental Engineering **2015**
University of California, Irvine *Irvine, California*

Master of Science in Civil and Environmental Engineering **2009**
University of Tehran *Tehran, Iran*

Bachelor of Science in Civil and Environmental Engineering **2007**
Ferdowsi University of Mashhad *Mashhad, Iran*

RESEARCH EXPERIENCE

Graduate Research Assistant **2010–2015**
University of California, Irvine *Irvine, California*

Graduate Research Assistant **2007–2009**
University of Tehran *Tehran, Iran*

Undergraduate Intern **2007**
Toosab Consulting Engineers Co. *Mashhad, Iran*

TEACHING EXPERIENCE

Instructor: Engineering Problem Solving **Summer 2012–2013**
Teaching Assistant: Engineering Problem Solving **2012–2014**
University of California, Irvine *Irvine, California*

Teaching Assistant: Strength of Material **Fall 2005**
Tutor: Design of Steel Structures **Spring 2007**
Tutor: Hydraulic Structures **Spring 2007**
Ferdowsi University of Mashhad *Mashhad, Iran*

HONORS AND AWARDS

The Henry Samueli Endowed fellowship (18,000 USD), University of California, Irvine, 2010.
Nominated for Teaching Assistant of the School of Engineering, University of California, Irvine, 2013-2014.
Top student in master's degree, University of Tehran, IRAN, 2007-2009.
Top 5% students in bachelor's degree, Ferdowsi University of Mashhad, IRAN, 2003-2007.

REFEREED JOURNAL PUBLICATIONS

J.A. Vrugt, and M. Sadegh, The scientific method, Bayes theorem, diagnostic model evaluation, epistemic errors, and summary metrics as prior distribution **2015**
Water Resources Research, Under Preparation

M. Sadegh, J.A. Vrugt, H.V. Gupta, The soil water characteristic as new class of closed-form parametric expressions for the flow duration curve **2015**
Water Resources Research, Under Review

M. Sadegh, J.A. Vrugt, C. Xu, E. Volpi, Diagnosing non-stationary watershed behavior from process-based model evaluation using approximate Bayesian computation **2015**
Water Resources Research, Under Review

T. Lochbühler, J.A. Vrugt, M. Sadegh, N. Linde, Summary statistics from training images as prior information in probabilistic inversion **2015**
Geophysical Journal International, In Press

M. Sadegh, J.A. Vrugt, Approximate Bayesian computation using Markov chain Monte Carlo simulation: DREAM$_{(\mathrm{ABC})}$ **2014**
Water Resources Research, 50, doi:10.1002/2014WR015386.

J. A. Vrugt, M. Sadegh, Towards diagnostic model evaluation: Approximate Bayesian computation **2013**
Water Resources Research, 49, 43354345, doi:10.1002/wrcr.20354.

M. Sadegh, J.A. Vrugt, Bridging the gap between GLUE and formal statistical approaches: approximate Bayesian computation **2013**
Hydrology and Earth System Sciences, 17, 4831-4850, doi:10.5194/hess-17-4831-2013.

M. Sadegh, R. Kerachian, Water resources allocation using solution concepts of fuzzy cooperative games: Fuzzy least core and Fuzzy weak least core **2011**
Water Resources Management, 25, 2543-2573, doi: 10.1007/s11269-011-9826-x.

M. Sadegh, N. Mahjouri and R. Kerachian, Optimal inter-basin water allocation using crisp and Fuzzy Shapley games **2010**
Water Resources Management, 24, 22912310, doi: 10.1007/s11269-009-9552-9.

REFEREED CONFERENCE PUBLICATIONS

M. Sadegh, N. Mahjouri, R. Kerachian, Developing a Fuzzy game theoretic model for optimal inter-basin water allocation **2010**
Proceedings of EWRIs 3rd developing nations conference: India

M. Sadegh, R. Kerachian, Fuzzy cooperative game solution concepts to a water resource allocation problem **2010**
Proceedings of EWRIs 3rd developing nations conference: India

TALKS AND LECTURES

J.A. Vrugt, M. Sadegh, Model data synthesis of complex CPU-intensive system models **2014**
AGU Fall meeting

J.A. Vrugt, M. Sadegh, A diagnostic approach to model evaluation: Approximate Bayesian computation **2014**
SIAM conference on uncertainty quantification

Approximate Bayesian computation for model evaluation in hydrology **2014**
The University of Arizona

M. Sadegh, J.A. Vrugt, Approximate Bayesian computation using Markov chain Monte Carlo simulation: Theory, concepts, and applications **2013**
AGU Fall meeting

M. Sadegh, J.A. Vrugt, Reconciling formal and informal statistical approaches: Approximate Bayesian computing and GLUE **2012**
AGU Fall meeting

SOFTWARE

MATLAB, R, C++

ABSTRACT OF THE DISSERTATION

Towards Improved Model Evaluation:
Diagnostic, Likelihood-free, Bayesian Inference

By

Mojtaba Sadegh

Doctor of Philosophy in Civil and Environmental Engineering

University of California, Irvine, 2015

Professor Jasper A. Vrugt, Chair

In the past decades, Bayesian methods have found widespread application and use in environmental systems modeling. Bayes theorem states that the posterior probability of a hypothesis is proportional to the product of the prior probability of this hypothesis and the likelihood of the hypothesis given the new/incoming observations. In science and engineering, this hypothesis often constitutes some numerical simulation model, which summarizes using algebraic, empirical, and differential equations, state variables and fluxes, all our theoretical and/or practical knowledge of the system of interest, and unknown model parameters which are subject to inference using some data of the observed system response. The Bayesian approach is intimately related to the scientific method and uses an iterative cycle of hypothesis formulation (model), experimentation and data collection, and theory/hypothesis refinement to elucidate the rules that govern the natural world. Unfortunately, model refinement has proven to be very difficult in large part because of the poor diagnostic power of residual based likelihood functions.

In this thesis, I present seven different chapters (publications) that introduce the theory, concepts and implementation of a diagnostic approach to model identification and evaluation. This approach, coined approximate Bayesian computation (ABC), relaxes the need

for an explicit likelihood function in favor of one or more summary statistics, which when rooted in the relevant environmental theory have a much stronger and compelling diagnostic power than some average measure of the size of the error residuals. The proposed methodology is statistically coherent, and provides new insights into and guidance on model structural (epistemic) errors, model (hypothesis) refinement, system nonstationarity, and the information content of experimental data. What is more, the proposed diagnostic approach provides a much-needed statistical underpinning of the popular generalized likelihood uncertainty framework (GLUE) of Beven and co-workers, and presents a powerful alternative to subjective regularized inversion methods used in geophysical inversion. Finally, the DREAM(ABC) algorithm, developed to solve the diagnostic model evaluation problem, is orders of magnitude more efficient than commonly used ABC sampling methods, thereby permitting inference of parameter-rich system models.

Chapter 1

Introduction

1.1 Traditional Model-Data Fusion

In a common inverse problem, we wish to estimate the parameters, $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_d\}$ of a model, $\mathcal{H}$, given observations of the system behavior, $\tilde{\mathbf{Y}} = \{\tilde{y}_1, \ldots, \tilde{y}_n\}$. The observations or data are linked to the unknown parameters $\boldsymbol{\theta}^*$ through some physical system, $\Im$

$$\tilde{\mathbf{Y}} \leftarrow \Im(\boldsymbol{\theta}^*) + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $\boldsymbol{\varepsilon} = \{\varepsilon_1, \ldots, \varepsilon_n\}$ is a $n \times 1$-vector of measurement errors. When a model hypothesis, or simulator, $\mathbf{Y} \leftarrow \mathcal{H}(\boldsymbol{\theta}^*, \tilde{\mathbf{u}}, \tilde{\mathbf{x}}_0)$ of the physical process is available, one can model the data

$$\tilde{\mathbf{Y}} \leftarrow \mathcal{H}(\boldsymbol{\theta}^*, \tilde{\mathbf{u}}, \tilde{\mathbf{x}}_0) + \mathbf{e}, \tag{1.2}$$

where $\tilde{\mathbf{u}} = \{\tilde{u}_1, \ldots, \tilde{u}_n\}$ denotes the forcing data, $\tilde{\mathbf{x}}_0$ signifies the initial states, and $\mathbf{e} = \{e_1, \ldots, e_n\}$ includes observation error (input and calibration data) as well as error due to

the fact that the simulator, $\mathcal{H}(\boldsymbol{\theta}^*|\cdot)$ may be systematically different from reality, $\Im(\boldsymbol{\theta}^*)$ for the parameters $\boldsymbol{\theta}^*$. The latter may arise from, e.g., numerical error, spatial discretization and improper (conceptual) model formulation [*Sadegh and Vrugt*, 2013].

Figure 1.1 provides an overview of possible error sources that affect our ability to correctly describe the physical system, $\Im(\boldsymbol{\theta}^*)$ of interest. Forcing data, model parameter, model state, and calibration data error are represented with a probability density function (pdf), whose statistical properties are typically unknown. Errors in the modeled (5) output, $y_{t(t>0)}$ and (6) state, $\mathbf{x}_{t(t>0)}$ dynamics originate from a wide variety of different error sources, including (1) inadequate and/or incomplete knowledge of the model parameters, $\boldsymbol{\theta}^*$ (2) errors in the input (forcing) data, $\mathbf{u}$ and (3) initial states, $\mathbf{x}_0$ (4) structural inadequacies in the model equations, and/or improper dimensionality of the state space, and (7) errors in the calibration data, $\tilde{y}_{t(t>0)}$. The mathematical operator $\bigotimes$ (also called "likelihood function") is used to judge the distance between the model predictions and corresponding calibration data. This function should explicitly recognize the contribution and role of each individual error source in determining the error residual, but is very difficult to specify correctly with very weak prior information, and hence the pitchfork symbol is used [*Sadegh and Vrugt*, 2013].

Within the context of hydrologic modeling, measured rainfall depths, and estimates of (potential) evapotranspiration typically constitute the main forcing data. These two input variables strongly determine the simulated streamflow at interior points and the catchment outlet, surface runoff, soil moisture fluxes and storage of water in the catchment. Examples of model states are soil moisture, groundwater table depth, and hydraulic heads (amongst others). Their knowledge is beneficial to adequately represent the storage of water in the variably saturated zone and groundwater, and hence ensure an adequate model calibration. Finally, calibration data often involves time series of (spatially-distributed) streamflow observations, or time-lapse measurements of tracer concentrations. Inevitably, each of these data sources is subject to uncertainty, which severely complicates parameter estimation and

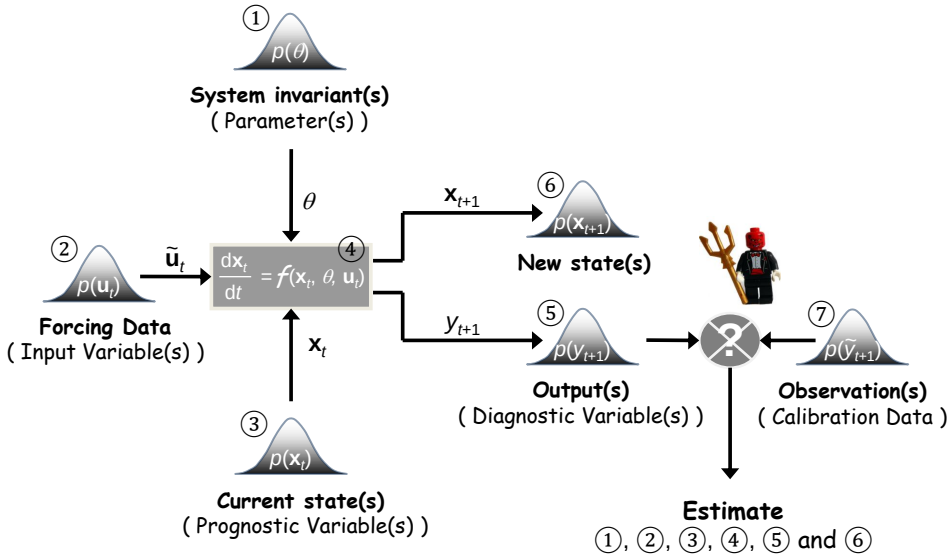


Figure 1.1: Explicit recognition of the role of 1: parameter, 2: forcing data, 3: initial state, 4: model structural, 5: output, 6: state, and 7: calibration data uncertainty. The pitchfork symbol illuminates the difficulty which formulation of the likelihood function (and prior distribution / parameterization of individual error sources) to use to summarize the error residuals. Explicit treatment of individual error sources is required to increase the prospects of explaining the reasons for model inadequacy and learning from the experimental data [*Sadegh and Vrugt*, 2013].

quantification of model structural errors [*Sadegh and Vrugt*, 2013].

During the past 4 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 [*Vrugt and Sadegh*, 2013].

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 (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,

$$F_{\mathrm{SLS}}(\theta) = \sum_{t=1}^{n} e_t^2(\theta) \tag{1.3}$$

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 [*Vrugt and Sadegh*, 2013].

Several contributions to the hydrologic literature have criticised 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 and forcing data errors, this will be purely for the sake of mathematical convenience [*Vrugt and Sadegh*, 2013].

Some interesting approaches for addressing the limitations of the classical calibration ap-

proach, 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 (BATEA) framework of Kavetski and coworkers [*Kavetski et al.*, 2006a,b, *Kuczera et al.*, 2006, *Thyer et al.*, 2009, *Renard et al.*, 2011a], the SODA, 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* [2010a] combined with Markov Chain Monte Carlo (MCMC) simulation using DREAM [*Vrugt et al.*, 2008a,b, 2009a, *Laloy and Vrugt*, 2012], (Bayesian) model-averaging [*Butts et al.*, 2004, *Ajami et al.*, 2007, *Vrugt and Robinson*, 2007a, *Rings et al.*, 2012], the hypothetico-inductive data-based mechanistic modeling framework of *Young* [2013] 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 1.1) [*Vrugt and Sadegh*, 2013].

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

$$p(\theta|\tilde{\mathbf{Y}}) = \frac{p(\theta)p(\tilde{\mathbf{Y}}|\theta)}{p(\tilde{\mathbf{Y}})}, \tag{1.4}$$

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 normalization constant or evidence, $p(\tilde{\mathbf{Y}})$ is required for Bayesian model selection and averaging [*Marshall et al.*, 2004, *Schoups and Vrugt*, 2011, *Laloy and Vrugt*, 2012], but if our interest is only in the parameters all our statistical inferences (mean, standard deviation, etc.) can be made from the unnormalized density [*Vrugt and Sadegh*,

2013]:

$$p(\theta|\tilde{\mathbf{Y}}) \propto p(\theta)L(\theta|\tilde{\mathbf{Y}}), \tag{1.5}$$

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

$$L(\theta|\tilde{\mathbf{Y}}) = \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\hat{\sigma}_\nu^2}} \exp\left[-\frac{1}{2}\hat{\sigma}_\nu^{-2}(\tilde{y}_t - y_t(\theta))^2\right], \tag{1.6}$$

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, *Bikowski et al.*, 2012, *Laloy and Vrugt*, 2012].

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 (1.6) have been proposed to handle nontraditional residual distributions [*Schoups and Vrugt*, 2010a, *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 average) properties of the error residuals - dilutes and mixes the available information into an index having little remaining correspondence to specific behaviours of the system. Whereas, this classical approach works for relatively simple (low order) models and allows for some treat-

ment 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, *Vrugt and Sadegh*, 2013].

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 system behaviors, and diagnostic evaluation proceeds with analysis of the behavioural (signature) similarities and differences between the system data and the model simulations. Ideally, correction then proceeds by relating these symptoms of model malfunction to individual process descriptions and model components. *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 [*Vrugt and Sadegh*, 2013].

1.2 Likelihood-Free Inference: Approximate Bayesian Computation

Vrugt and Sadegh [2013] advocated the use of "likelihood-free" inference for diagnostic model evaluation. This approach, introduced in the statistical literature about three decades ago [*Diggle and Gratton*, 1984], is especially useful for cases where the likelihood is intractable, too expensive to be evaluated, or impossible to be formulated explicitly. This class of methods is also referred to as Approximate Bayesian Computation (ABC), a term coined by *Beaumont*

et al. [2002], and widens the realm of models for which statistical inference can be considered [*Marjoram et al.*, 2003, *Sisson et al.*, 2007, *Del Moral et al.*, 2008, *Joyce and Marjoram*, 2008, *Grelaud et al.*, 2009, *Ratmann et al.*, 2009]. ABC has rapidly gained popularity in the past few years, in particular for the analysis of complex problems arising in population genetics, ecology, epidemiology, and systems biology.

The premise behind ABC is that $\boldsymbol{\theta}^*$ should be a sample from the posterior distribution if the distance between the observed and simulated data, $\rho\big(\tilde{\mathbf{Y}}, \mathbf{Y}(\boldsymbol{\theta}^*)\big)$, is smaller than some small positive value, ϵ [*Marjoram et al.*, 2003, *Sisson et al.*, 2007]. Figure 1.3 provides a conceptual overview of the ABC methodology. All ABC based methods approximate the likelihood function by simulations, the outcomes of which are compared with the observed data [*Beaumont*, 2010, *Bertorelle et al.*, 2010, *Csilléry et al.*, 2010]. In so doing, ABC algorithms attempt to approximate the posterior distribution by sampling from

$$p(\boldsymbol{\theta}|\tilde{\mathbf{Y}}) \propto \int_{\boldsymbol{\mathcal{Y}}} p(\boldsymbol{\theta})\text{Model}(\boldsymbol{y}|\boldsymbol{\theta})\text{I}(\rho\big(\tilde{\mathbf{Y}}, \mathbf{Y}(\boldsymbol{\theta})\big) \leq \epsilon)d\boldsymbol{y}, \tag{1.7}$$

where $\boldsymbol{\mathcal{Y}}$ denotes the support of the simulated data, $\mathbf{Y} \sim \text{Model}(\boldsymbol{y}|\boldsymbol{\theta})$, and $\text{I}(a)$ is an indicator function that returns one if the condition a is satisfied and zero otherwise. The accuracy of the estimated posterior distribution, $p(\theta|\rho\big(\tilde{\mathbf{Y}}, \mathbf{Y}(\boldsymbol{\theta})\big) \leq \epsilon)$ depends on the value of ϵ. In the limit of $\epsilon \to 0$ the sampled distribution will converge to the true posterior, $p(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$ [*Pritchard et al.*, 1999, *Beaumont et al.*, 2002, *Ratmann et al.*, 2009, *Turner and van Zandt*, 2012]. Yet, this requires the underlying model operator to be stochastic, and hence Model$(\cdot)$ in Equation (1.7) is equivalent to the output of the deterministic model $\mathcal{H}$ in Equation (1.2) plus a random error with probabilistic properties equal to those of $\mathbf{e}$ [*Vrugt and Sadegh*, 2013].
A pseudo-code of the generic ABC approach is given below.

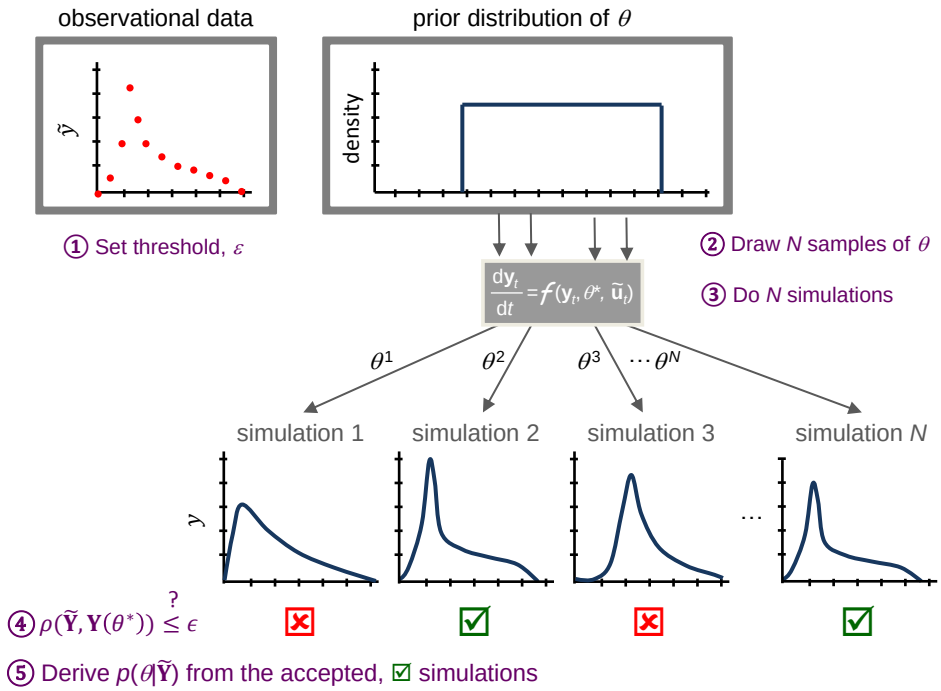


Figure 1.2: Conceptual overview of approximate Bayesian computation (ABC) for a hypothetical one-dimensional parameter estimation problem. First, N samples are drawn from a user-defined prior distribution, $\theta^* \sim p(\theta)$. Then, this ensemble is evaluated with the model and creates N model simulations. If the distance between the observed and simulated data, $\rho(\tilde{\mathbf{Y}}, \mathbf{Y}(\theta^*))$ is smaller than or equal to some nominal value, ϵ then θ^* is retained, otherwise the simulation is discarded. The accepted samples are then used to approximate the posterior parameter distribution, $p(\theta|\tilde{\mathbf{Y}})$. Note that for sufficiently complex models and large data sets the probability of happening upon a simulation run that yields precisely the same simulated values as the observations will be very small, often unacceptably so. Therefore, $\rho(\tilde{\mathbf{Y}}, \mathbf{Y}(\boldsymbol{\theta}^*))$ is typically defined as a distance between summary statistics of the simulated, $S(\mathbf{Y}(\boldsymbol{\theta}^*))$ and observed, $S(\tilde{\mathbf{Y}})$ data, respectively. Modified after *Sunnåker et al.* [2013].

Algorithm 1 Rejection sampler

for $i = 1, \ldots, N$ **do**

  repeat

    generate θ^* from the prior distribution, $\theta^* \sim p(\theta)$

    simulate $\mathbf{Y}$ using the model, $\mathbf{Y} \leftarrow \mathcal{H}(\theta^*|\cdot)$

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

  set $\theta_i \leftarrow \theta^*$

  set $w_i \leftarrow \frac{1}{N}$

end for

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, $\mathbf{Y} \leftarrow \mathcal{H}(\theta^*|\cdot)$. 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 pre-defined 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, $\hat{p}(\theta|\rho(\tilde{\mathbf{Y}}, \mathbf{Y}) \leq \epsilon)$. This algorithm converges to the true posterior $p(\theta|\tilde{\mathbf{Y}})$ when $\epsilon \rightarrow 0$, provided that the summary statistic(s), $S(\cdot)$ is (are) near sufficient [*Pritchard et al.*, 1999, *Beaumont et al.*, 2002, *Ratmann et al.*, 2009, *Turner and van Zandt*, 2012]. Note that for very small values of ϵ, the rejection rates can be dramatically high, and Algorithm 1 thus very inefficient [*Vrugt and Sadegh*, 2013].

For sufficiently complex system models and/or large data sets, it will be difficult, if not impossible, to find a model simulation that always fits the data within ϵ. It is often convenient to define $\rho(\tilde{\mathbf{Y}}, \mathbf{Y})$ as a distance between summary statistics, $S(\mathbf{Y})$ and $S(\tilde{\mathbf{Y}})$ of the simulated and observed data, respectively. For example, $S(\cdot)$ could be the sample mean, $\rho(\tilde{\mathbf{Y}}, \mathbf{Y}) = |\bar{\tilde{\mathbf{Y}}} - \bar{\mathbf{Y}}|$. 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 data [*Ratmann et al.*, 2009]. Such approach has many elements in common with multi-criteria analysis, but benefits from a proper statistical foundation [*Vrugt and Sadegh*, 2013].

1.3 Example

To illustrate the ABC methodology, lets consider a simple linear regression model, $y_t = \theta_1 t + \theta_2$ as 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 1.3 plots the actual data (solid circles), and the original model (solid black line).

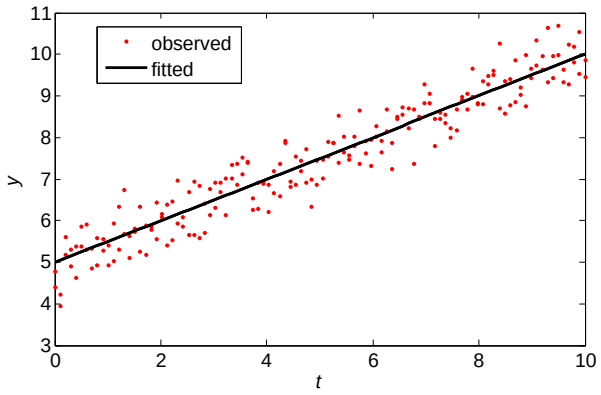


Figure 1.3: Synthetic data (solid circles) of the toy example and the original linear regression function, $Y \leftarrow \frac{1}{2}t + 5, t \in \{0, 10\}$ (solid black line) that was used to create the observations. The ABC framework now compares summary statistics of the measurement and simulated data to estimate the slope and intercept of the regression function.

The top panel in Figure 1.4 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 [*Vrugt and Sadegh*, 2013].

We now use ABC with summary statistic the mean of the actual data, $\rho(\tilde{\mathbf{Y}}, \mathbf{Y}(\theta)) = |\bar{\tilde{\mathbf{Y}}} - \bar{\mathbf{Y}}(\theta)|$. A uniform prior with $\theta_1 \in [0, 1]$ and $\theta_2 \in [-10, 10]$ was used in all our calculations. The results of ABC analysis are presented in the second panel (from top) in Figure 1.4 and

correspond to $\epsilon = 0.005$. The marginal distributions settle around their "true" values, but demonstrate significant dispersion, much larger than what can be expected from the data. From the outset 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 of the simulated data similar to that of the observations. We therefore add, as second metric, the standard deviation of the data (observations), $\text{std}(\tilde{\mathbf{Y}})$ using

$$\rho(\tilde{\mathbf{Y}}, \mathbf{Y}) = \max\left(\left|\bar{\tilde{\mathbf{Y}}} - \bar{\mathbf{Y}}(\theta^*)\right|, \left|\text{std}(\tilde{\mathbf{Y}}) - \text{std}(\mathbf{Y}(\theta^*))\right|\right), \tag{1.8}$$

as a distance function. The bottom panel of Figure 1.4 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 diagnostics we derive similar posterior histograms as those derived from a classical first-order statistical analysis [*Vrugt and Sadegh*, 2013].

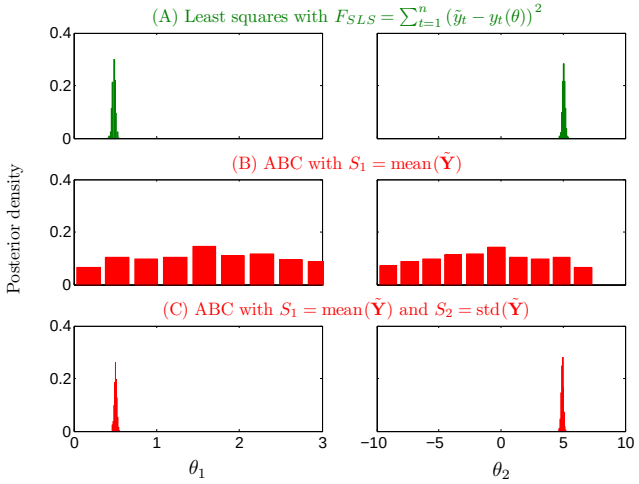


Figure 1.4: Marginal posterior distribution of the slope, θ_1 and intercept, θ_2 derived from least squares fitting (top panel) and ABC inference using one (mean of data) (middle panel) and two (mean and standard deviation of data (bottom panel) summary diagnostic(s).

1.4 Summary statistics versus residual-based likelihood functions

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. This relaxes the need for the residual-based likelihood functions, the very construction of which dilutes and mixes the available information into an index having little remaining correspondence to specific behaviours 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]. The advantage of summary statistics is that they exhibit a much better diagnostic power than some convoluted set of error residuals, and if properly designed and rooted in the relevant environmental theory, can help to illuminate to what degree a representation of the real world has been adequately achieved and how the model (null hypothesis) should be refined. The ABC methodology is particularly useful for diagnostic model calibration and evaluation. By defining summary metrics so that they each measure different parts of the system behavior, the posterior distribution could help to pin point which individual components of the model are in need of improvement [*Vrugt and Sadegh*, 2013].

1.5 Scope

The 2nd chapter of this thesis introduces the ABC framework for diagnostic model analysis and evaluation. The ever increasing pace of computational power, along with continued advances in measurement technologies and improvements in process understanding has stim-

ulated 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) towards 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 chapter, we introduce 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 more clear and 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 [*Vrugt and Sadegh*, 2013].

In the 3rd chapter this thesis, we will demonstrate the equifinality of the proposed ABC framework with the commonly used Generalized Likelihood Uncertainty Estimation (GLUE) framework of Beven and coworkers. In recent years, a strong debate has emerged in the hydrologic literature how to properly treat nontraditional error residual distributions and quantify parameter and predictive uncertainty. Particularly, there is strong disagreement whether such uncertainty framework should have its roots within a proper statistical (Bayesian) context using Markov chain Monte Carlo (MCMC) simulation techniques, or whether such a framework should be based on a quite different philosophy and implement informal likelihood functions and simplistic search methods to summarize parameter and predictive distribu-

tions. This chapter demonstrates that Approximate Bayesian Computation (ABC) bridges the gap between formal and informal statistical model-data fitting approaches. This work further studies the theoretical and numerical equivalence of formal and informal (GLUE) Bayesian approaches using discharge and forcing data from different watersheds in the United States. We demonstrate that the limits of acceptability approach of GLUE is a special variant of ABC if each discharge observation of the calibration data set is used as a summary diagnostic [*Sadegh and Vrugt*, 2013].

In the 4th chapter of this thesis we propose several methodological improvements over commonly used ABC sampling methods to permit inference of complex system models. Our methodology entitled, DREAM$_{(\mathrm{ABC})}$ uses the DiffeRential Evolution Adaptive Metropolis algorithm [*Vrugt et al.*, 2008a, 2009a] as its main building block and takes advantage of a continuous fitness function to efficiently explore the behavioral model space. Three case studies demonstrate that DREAM$_{(\mathrm{ABC})}$ is at least 3 - 1,000 times more efficient than commonly used ABC sampling methods. DREAM$_{(\mathrm{ABC})}$ is also more amendable to distributed, multiprocessor implementation, a prerequisite to diagnostic inference of CPU-intensive system models [*Sadegh and Vrugt*, 2014].

The 5th chapter of this thesis demonstrates the diagnostic power of the proposed ABC methodology to detect nonstationary behavior of watersheds. A key assumption in virtually all modeling studies in surface hydrology is that the watershed properties (dominant flowpaths, soil and vegetation properties, etc.) are time-invariant, and hence can be estimated from a sufficiently long rainfall-discharge data record. However, this stationarity assumption has been criticized in the recent literature. Human interference and climate change can substantially affect land-use type and distribution, vegetation and soil properties (among others), and thus possibly render the runoff response to rainfall nonstationary. In this chapter we will introduce a process-based methodology that will help to assess catchment nonstationarity. Our approach has its roots in diagnostic model evaluation, and uses approximate

Bayesian computation (ABC) and Markov chain Monte Carlo simulation with DREAM$_{(ABC)}$ to detect gradual/abrupt changes in the watersheds response to rainfall. Nonstationarity of catchment behavior is not readily apparent from statistical analysis and/or visual interpretation of the observed system response, but is evident if summary metrics are calculated for different time periods. Models that assume stationarity are unable to accurately simulate these gradual/abrupt shifts in catchment behavior, and consequently, diagnostic inference with ABC cannot produce a proper behavioral (posterior) model space. Our preliminary studies demonstrate that the presented methodology is able to differentiate successfully between watersheds that are classified as stationary and those that have undergone significant changes in land-use, urbanization and/or experienced a significant characteristic change in rainfall forcing, and thus are deemed nonstationary [*Sadegh et al.*, 2015a].

The 6th chapter of this thesis is focused on the flow duration curve as a signature catchment characteristic that depicts graphically the relationship between the exceedance probability of streamflow and its magnitude. This curve is relatively easy to create and interpret, and is used widely for hydrologic analysis, water quality management, and the design of hydroelectric power plants (among others). Several mathematical formulations have been proposed to mimic the FDC. Yet, these efforts have not been particularly successful, in large part because available functions are not flexible enough to portray accurately the functional shape of the FDC for a large range of different catchments and hydrologic behaviors. Here, we draw inspiration from soil physics and propose commonly used models of the soil water characteristic as new class of closed-form parametric formulations for the flow duration curve. These soil water retention functions are relatively simple, contain between 2 to 5 fitting coefficients (parameters), and describes closely the observed FDCs of the 438 catchments of the MOPEX data set. If quality of fit is of main importance then the 5-parameter bimodal retention function of Kosugi is preferred, whereas a 2-parameter formulation of this model is deemed superior for diagnostic analysis as the fitting coefficients exhibit a negligible correlation [*Sadegh et al.*, 2015b].

Application of the ABC framework in the field of geophysical modeling is presented in chapter 7 of this thesis. In this chapter, a strategy is presented to incorporate prior information from conceptual geological models in probabilistic inversion of geophysical data. The geological models are represented by multiple-point statistics training images featuring the expected lithological units and structural patterns. Information from an ensemble of training image realizations is used in two different ways. First, dominant modes are identified by analysis of the frequency content in the realizations, which drastically reduces the model parameter space in the frequency-amplitude domain. Second, the distributions of global, summary metrics such as model roughness, parameter variability and percentile distribution of subsurface properties are used to formulate a prior probability density function (pdf). The inverse problem is formulated in a Bayesian framework and the posterior pdf is sampled using Markov chain Monte Carlo simulation. The usefulness and applicability of this method is demonstrated on two case studies in which synthetic crosshole ground-penetrating radar travel time data are inverted to recover 2-D porosity fields. The use of prior information from training images significantly enhances the reliability of the posterior models by removing inversion artifacts and improving individual parameter estimates. The proposed methodology reduces the ambiguity inherent in the inversion of high-dimensional parameter spaces, accommodates a wide range of summary statistics and geophysical forward problems [*Lochbühler et al.*, 2014c].

Chapter 2

Towards Diagnostic Model Calibration and Evaluation: Approximate Bayesian Computation

This Chapter has been published in Water Resources Research. Citation: Vrugt, J.A., and M. Sadegh (2013), Toward diagnostic model calibration and evaluation: Approximate Bayesian computation, *Water Resources Research*, *49*, doi:10.1002/wrcr.20354.

2.1 Introduction and scope

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,

$$\mathbf{Y} \leftarrow f(\mathbf{x}_0, \theta, \tilde{\mathbf{U}}), \tag{2.1}$$

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,

$$\mathbf{E}(\theta) = \digamma[\tilde{\mathbf{Y}}] - \digamma[\mathbf{Y}(\theta)] = \{e_1(\theta), \ldots, e_n(\theta)\}, \tag{2.2}$$

where $\digamma[\cdot]$ allows for linear (nonlinear) 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.

During the past 4 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.

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 (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,

$$F_{\mathrm{SLS}}(\theta) = \sum_{t=1}^{n} e_t^2(\theta) \tag{2.3}$$

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.

Several contributions to the hydrologic literature have criticised 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 and forcing data errors, this will be purely for the sake of mathematical convenience.

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 (BATEA) framework of Kavetski and coworkers [*Kavetski et al.*, 2006a,b, *Kuczera et al.*, 2006, *Thyer et al.*, 2009, *Renard et al.*, 2011a], the SODA, 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* [2010a] combined with Markov Chain Monte Carlo (MCMC) simulation using DREAM [*Vrugt et al.*, 2008a,b, 2009a, *Laloy and Vrugt*, 2012], (Bayesian) model-averaging [*Butts et al.*, 2004, *Ajami et al.*, 2007, *Vrugt and Robinson*, 2007a, *Rings et al.*, 2012], the hypothetico-inductive data-based mechanistic modeling framework of *Young* [2013] 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 2.1).

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

$$p(\theta|\tilde{\mathbf{Y}}) = \frac{p(\theta)p(\tilde{\mathbf{Y}}|\theta)}{p(\tilde{\mathbf{Y}})}, \tag{2.4}$$

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 normalization constant or evidence, $p(\tilde{\mathbf{Y}})$ is required for Bayesian model selection and averaging [*Marshall et al.*, 2004, *Schoups and Vrugt*, 2011, *Laloy and Vrugt*, 2012], but if our interest is only in the parameters all our statistical inferences (mean, standard deviation, etc.) can be made from the unnormalized density:

$$p(\theta|\tilde{\mathbf{Y}}) \propto p(\theta)L(\theta|\tilde{\mathbf{Y}}), \tag{2.5}$$

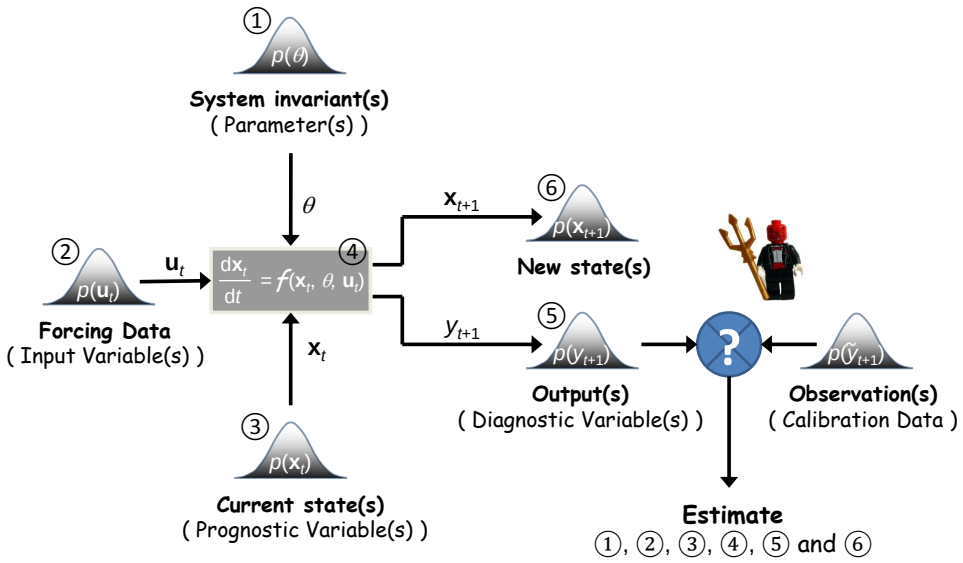


Figure 2.1: Bayesian model evaluation framework that explicitly recognizes the role of (1) parameter, (2) forcing data, (3), initial state, (4) model structural, (5) output, and (6) state uncertainty. The symbol $\bigotimes$ illuminates the difficulty which likelihood function (and prior distribution / 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.

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

$$L(\theta|\tilde{\mathbf{Y}}) = \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\hat{\sigma}_\nu^2}} \exp\left[-\frac{1}{2}\hat{\sigma}_\nu^{-2}(\tilde{y}_t - y_t(\theta))^2\right], \tag{2.6}$$

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, *Bikowski et al.*, 2012, *Laloy and Vrugt*, 2012].

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 $L(\theta|\tilde{\mathbf{Y}})$

$$\ell(\theta|\tilde{\mathbf{Y}}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{\sigma}_\nu^2) - \frac{1}{2}\hat{\sigma}_\nu^{-2}\sum_{t=1}^{n}(\tilde{y}_t - y_t(\theta))^2, \tag{2.7}$$

which is equivalent to

$$\ell(\theta|\tilde{\mathbf{Y}}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{\sigma}_\nu^2) - \frac{1}{2}\hat{\sigma}_\nu^{-2}F_{\mathrm{SLS}}(\theta) \tag{2.8}$$

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.

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 (2.7) have been proposed to handle nontraditional residual distributions [*Schoups and Vrugt*, 2010a, *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 average) properties of the error residuals - dilutes and mixes the available information into an index having little remaining correspondence to specific behaviours 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].

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 system behaviors, and diagnostic evaluation proceeds with analysis of the behavioural (signature) similarities and differences between the system data and the model simulations. Ideally, correction then proceeds by relating these symptoms of model malfunction to individual process descriptions and model components. *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.

In this chapter 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.

2.2 Approximate Bayesian Computation

This chapter 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 espe-

cially 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 Approximate Bayesian Computation (ABC) and is currently a "hot" topic in statistics [*Marjoram et al.*, 2003, *Sisson et al.*, 2007, *Del Moral et al.*, 2008, *Joyce and Marjoram*, 2008, *Grelaud et 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 value, ϵ. So rather than working directly on the error residuals with an explicit likelihood function, ABC compares summery statistics of the calibration data and corresponding simulation [*Marjoram et al.*, 2003, *Sisson et al.*, 2007]

$$\rho(\tilde{\mathbf{Y}}, \mathbf{Y}(\theta^*)) \leq \epsilon, \tag{2.9}$$

to decide whether the model is close enough to the data and retain (accept) θ^*.

A pseudo-code of the generic ABC approach is given below.

Algorithm 2 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

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 pre-defined 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, $\hat{p}(\theta|\rho(\tilde{\mathbf{Y}}, \mathbf{Y}) \leq \epsilon)$. This algorithm converges to the true posterior $p(\theta|\tilde{\mathbf{Y}})$ when $\epsilon \to 0$, provided that the summary statistic(s), $S(\cdot)$ is (are) near sufficient [*Pritchard et al.*, 1999, *Beaumont et al.*, 2002, *Ratmann et al.*, 2009, *Turner and van Zandt*, 2012]. Note that for very small values of ϵ, the rejection rates can be dramatically high, and Algorithm 2 thus very inefficient.

For diagnostic model evaluation (or computational efficiency in case of large data sets), it is often convenient to define $\rho(\tilde{\mathbf{Y}}, \mathbf{Y})$ as a distance between summary statistics, $S(\mathbf{Y})$ and $S(\tilde{\mathbf{Y}})$ of the simulated and observed data, respectively. For example, $S(\cdot)$ could be the sample mean, $\rho(\tilde{\mathbf{Y}}, \mathbf{Y}) = |\bar{\tilde{\mathbf{Y}}} - \bar{\mathbf{Y}}|$. 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 data [*Ratmann et al.*, 2009]. Such approach has many elements in common with multi-criteria analysis, but benefits from a proper statistical foundation.

2.3 Case Study I: Toy Problem

To illustrate the ABC methodology, lets consider a simple linear regression model, $y_t = \theta_1 t + \theta_2$ as 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 2.2 plots the actual data (solid circles), and the original model (solid black line).

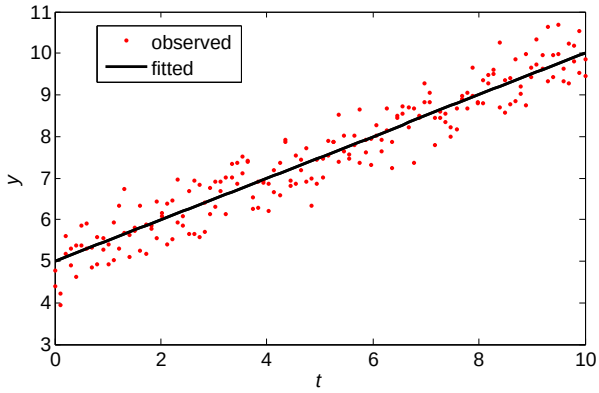


Figure 2.2: Synthetic data (solid circles) of the toy example and the original linear regression function, $Y \leftarrow \frac{1}{2}t + 5, t \in \{0, 10\}$ (solid black line) that was used to create the observations. The ABC framework now compares summary statistics of the measurement and simulated data to estimate the slope and intercept of the regression function.

The top panel in Figure 2.3 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.

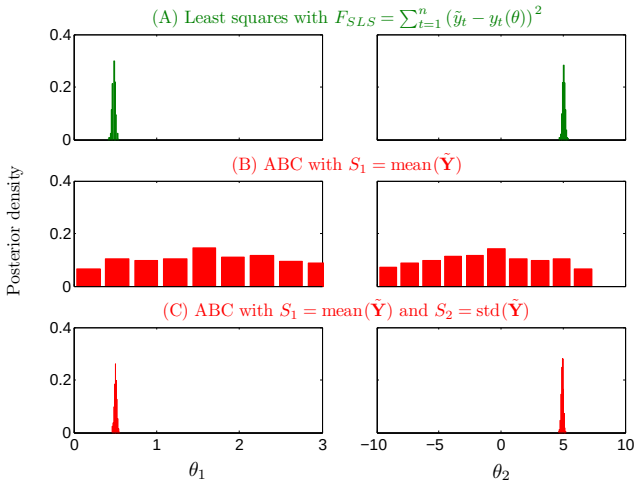


Figure 2.3: Marginal posterior distribution of the slope, θ_1 and intercept, θ_2 derived from least squares fitting (top panel) and ABC inference using one (mean of data) (middle panel) and two (mean and standard deviation of data (bottom panel) summary diagnostic(s).

We now use ABC with summary statistic the mean of the actual data, $\rho(\tilde{\mathbf{Y}}, \mathbf{Y}(\theta)) =$

$|\bar{\tilde{\mathbf{Y}}} - \bar{\mathbf{Y}}(\theta)|$. 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 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 the sampling distribution, $q_j(\theta_k, \cdot) = N_d(\theta_k, \sum); j > 1$ adapted using $\sum = \mathrm{Cov}(\theta_{1,j-1}, \ldots, \theta_{N,j-1})$ with selection probability of θ_k equal to its previous posterior weight, $w_{k,j-1}; k \in \{1, \ldots, N\}$. 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.

The results of ABC-PMC are presented in the second panel (from top) in Figure 2.3 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 outset 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 of the simulated data similar to that of the observations. We therefore add, as second metric, the standard deviation of the data (observations), $\mathrm{std}(\tilde{\mathbf{Y}})$ using

$$\rho(\tilde{\mathbf{Y}}, \mathbf{Y}) = \max\left(\left|\bar{\tilde{\mathbf{Y}}} - \bar{\mathbf{Y}}(\theta^*)\right|, \left|\mathrm{std}(\tilde{\mathbf{Y}}) - \mathrm{std}(\mathbf{Y}(\theta^*))\right|\right), \tag{2.10}$$

as a distance function. The bottom panel of Figure 2.3 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 diagnostics

we derive similar posterior histograms as those derived from a classical first-order statistical analysis.

We now analyze the effect of the choice of ϵ on the ABC results. Figure 2.4 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 $\epsilon = 0.1$ seems adequate. Lower values of ϵ provide similar posterior estimates, yet unnecessarily increase the computational burden of the ABC analysis.

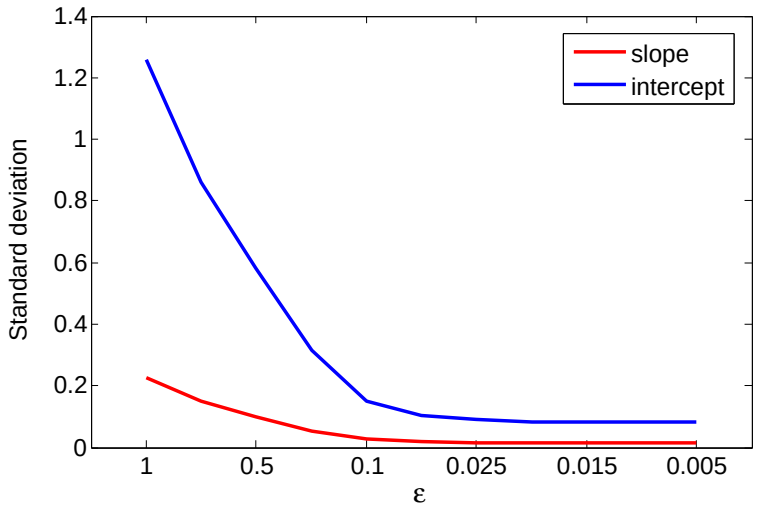


Figure 2.4: Relationship between the value of ϵ used in ABC-PMC and the posterior standard deviation of the slope (red line) and intercept (blue line). Results confirm theory that if $\epsilon \to 0$ the posterior statistics converge to their target values. A value of $\epsilon = 0.1$ generates stable results.

Table 2.1: Prior Uncertainty Ranges of Hydrologic Model Parameters.

Parameter	Symbol	Minimum	Maximum	Units
Maximum 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

2.4 Case Study II: Conceptual Hydrologic Modeling

We now move on to a more realistic 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 et al.* [2010b], and part of the DREAM software package. The model transforms rainfall into runoff at the watershed outlet using explicit process descriptions of interception, throughfall, evaporation, runoff generation, percolation, and surface and subsurface routing. Table 2.1 summarizes the seven different model parameters, and their prior uncertainty ranges.

Figure 2.5 presents the results of the ABC analysis using the mean and standard deviation of the discharge data as summary statistics. The top panel (Figure 2.5A) presents time series plots of the 95% posterior prediction uncertainty ranges (due to parameter uncertainty) and corresponding observations (solid circles). The bottom panel shows histograms of the marginal posterior parameter distributions for four randomly selected model parameters (Figure 2.5B-E), and the posterior RMSE values (Figure 2.5F). The streamflow prediction uncertainty intervals nicely encapsulate the observed data with a coverage of about 88%. The

parameters appear poorly defined with posterior ranges that extend the entire prior defined (uniform) hypercube. This result is perhaps not surprising, as the mean and standard deviation seem insufficient summary metrics to adequately constrain the hydrologic model parameters. The Root Mean Square Error (RMSE) values exhibit a log-normal distribution with values ranging between 0.67 and 2.04 mm/day. These values (and range) are much larger than what can be expected from a traditional explicit likelihood function that acts directly on the error residuals (more about this later).

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 baseflow index [*Gupta et al.*, 2008, *Yilmaz et al.*, 2008], and the flow duration curve [*Searcy*, 1959, *Westerberg et al.*, 2011]. The annual runoff coefficient, S_1 and annual baseflow index, S_2 are respectively defined as

$$S_1(\tilde{\mathbf{Y}}) = \frac{\sum_{t=1}^{365} \tilde{y}_t}{\sum_{t=1}^{365} \tilde{p}_t}, \qquad S_2(\tilde{\mathbf{Y}}) = \frac{\sum_{t=1}^{365} \tilde{y}_{b,t}}{\sum_{t=1}^{365} \tilde{y}_t}, \tag{2.11}$$

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

$$\tilde{y}_{b,t} = \phi \tilde{y}_{b,t-1} + \frac{1}{2}(1-\phi)(\tilde{y}_t + \tilde{y}_{t-1}), \tag{2.12}$$

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

Now we have defined an explicit measure of the annual runoff and baseflow index, we are now left with a summary metric for the flow duration curve. Unfortunately, it will be particularly difficult, if not impossible to adequately characterize the functional shape of the FDC with a single diagnostic measure. We therefore adopt a different approach and fit the following

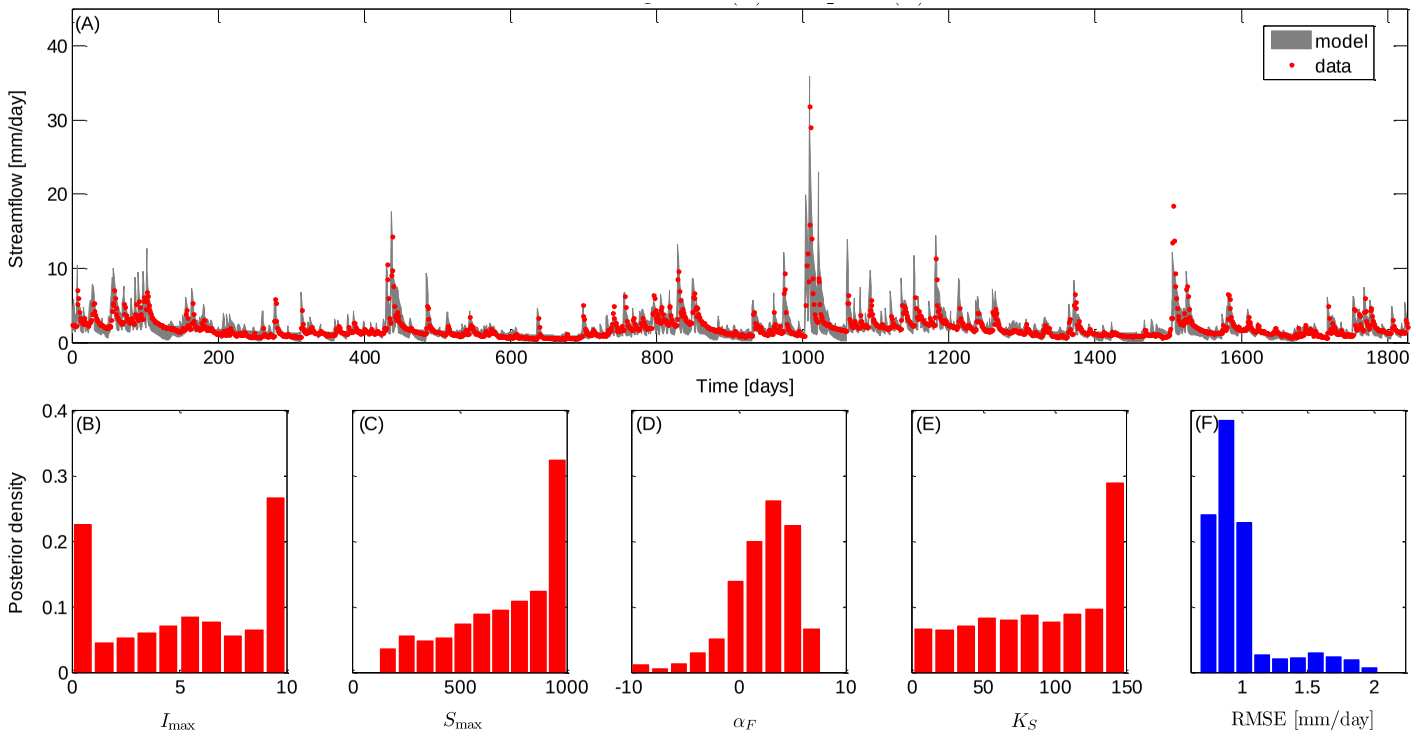


Figure 2.5: ABC inference for the French Broad River basin using the mean and standard deviation of the discharge data as summary statistics : (a) 95% posterior prediction uncertainty ranges (gray region) and corresponding discharge observations (solid circles), (b-e) histograms of the marginal posterior 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.

empirical model to the FDC,

$$F_t \leftarrow [1 + (S_3 \tilde{y}_t)^{S_4}]^{\left(\frac{1}{S_4} - 1\right)}, \tag{2.13}$$

where F_t denotes the exceedance probability of the t^{th} streamflow observation, and S_3 (day/mm) and S_4 (-) are calibration coefficients that serve as our third and fourth metric. This leaves us with four summary statistics for three different hydrologic signatures.

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

$$\rho(\tilde{\mathbf{Y}}, \mathbf{Y}(\theta^*)) = \max\left(\left|S_k(\tilde{\mathbf{Y}}) - S_k(\mathbf{Y}(\theta^*))\right|; k = \{1, \ldots, 4\}\right), \tag{2.14}$$

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.

Note that Equation (2.13) is equivalent to the water retention model of *van Genuchten* [1980] but with $\theta_s = 1$ and $\theta_s = 0$ so that $0 \leq F_t \leq 1$. Here we limit our attention to the French Broad River basin, and plot in Figure 2.6 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 0.01 in exceedance probability space and optimized values of $S_3 = 0.74$ and $S_4 = 3.34$. Chapter 6 of this thesis will study in depth the use of soil water retention functions as functional expressions of the flow duration curves.

Figure 2.7 presents the results of ABC using hydrologic summary statistics, $S_1 \rightarrow S_4$. The posterior parameter ranges (Figure 2.7B-E) have reduced somewhat compared to Figure 2.5, 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 baseflow index, and distribution of flow levels but with simulated streamflow response that

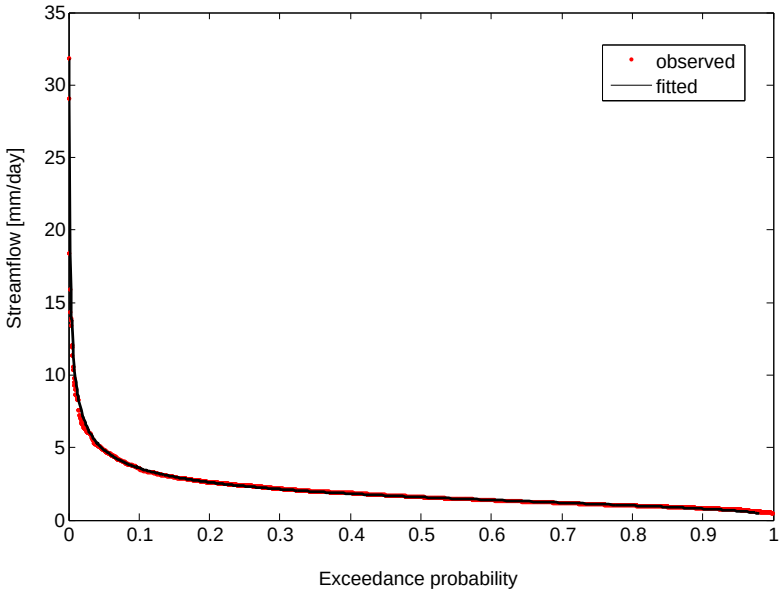


Figure 2.6: Observed (red dots) and model fitted (solid black line) flow duration curve (FDC) of the French Broad River basin at Asheville, North Carolina, USA. The two-parameter model closely matches the observed FDC. The optimized parameters, S_3 and S_4 of the FDC model are used as hydrologic signatures in our ABC analysis.

systematically underestimates the actual watersheds response to rainfall (Figure 2.7A). The recession periods appear to be well resolved by the model, but the peak flows are consistently underestimated. This deviation is so systematic that it is unlikely to be fully explained by errors in the discharge and precipitation data.

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 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 non-driven 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

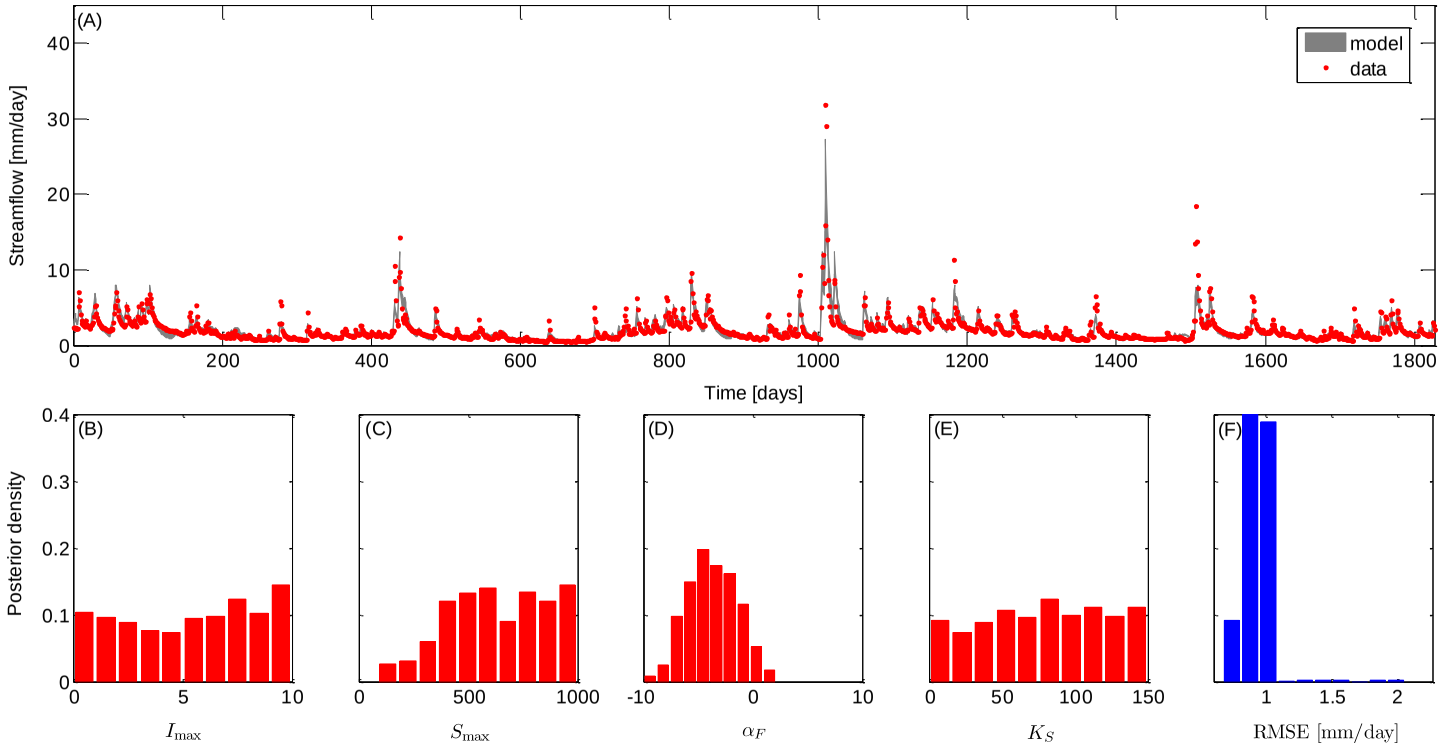


Figure 2.7: ABC inference for the French Broad River basin using summary metrics $S_1 \rightarrow S_4$ that represent the observed annual runoff coefficient, annual baseflow index, and flow duration curve: (a) 95% posterior prediction uncertainty ranges (gray region) and corresponding discharge observations (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 non-driven part of the hydrograph.

presence of preferential flow that is not accounted for by the model.

Note that the classical model calibration approach would not provide such guidance. For instance, consider the (least squares) likelihood function presented in Equation (2.8) 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/day) than the ABC-signature based analysis (Figure 2.7F) (0.67 - 2 mm/day 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 Figure 6 of *Schoups et al.* [2010b]). 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.

This example nicely illustrates the ability of ABC as paradigm for diagnostic model evaluation. We conclude this chapter with a short discussion.

2.5 Discussion

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 chapter in mind, three interesting ideas that immediately follow include

1. The impact of precipitation data uncertainty on the calibrated parameter estimates can be drastically reduced by using (inventing) 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 et al.*, 2010b]). 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.

2. There appears to be a close (theoretical) connection between GLUE and ABC. 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.

3. The ABC methodology provides a more rigorous statistical footing for multi-criteria methods [*Gupta et al.*, 1998] that use different parts of the model response or multiple different data types for calibration purposes.

2.6 Summary and Conclusions

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.

In this chapter we introduced Approximate Bayesian Computation 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 intro-

duce and illustrate the ABC framework. 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.

2.A Appendix A

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 3 ABC Population Monte Carlo sampler

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

Chapter 3

Approximate Bayesian Computation: Bridging the Gap Between Formal and Informal Statistical Approaches

This Chapter has been published in Hydrology and Earth System Sciences. Citation: Sadegh, M., and J.A. Vrugt (2013), Bridging the gap between GLUE and formal statistical approaches: approximate Bayesian computation, *Hydrology and Earth System Sciences*, *17*, 4831-4850, doi:10.5194/hess-17-4831-2013.

3.1 Introduction

In a common inverse problem, we wish to estimate the parameters, $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_d\}$ of a model, $\mathcal{H}$, given observations of the system behavior, $\tilde{\mathbf{Y}} = \{\tilde{y}_1, \ldots, \tilde{y}_n\}$. The observations or data are linked to the unknown parameters $\boldsymbol{\theta}^*$ through some physical system, $\Im$

$$\tilde{\mathbf{Y}} \leftarrow \Im(\boldsymbol{\theta}^*) + \boldsymbol{\varepsilon}, \tag{3.1}$$

where $\boldsymbol{\varepsilon} = \{\varepsilon_1, \ldots, \varepsilon_n\}$ is a $n \times 1$-vector of measurement errors. Examples of such problems are widespread in many different fields of study including medical imaging [*Kaipio et al.*, 2004], reservoir characterization [*Stenerud et al.*, 2008] and cosmology [*Jimenez et al.*, 2004]. When a model hypothesis, or simulator, $\mathbf{Y} \leftarrow \mathcal{H}(\boldsymbol{\theta}^*, \tilde{\mathbf{u}}, \tilde{\mathbf{x}}_0)$ of the physical process is available, one can model the data

$$\tilde{\mathbf{Y}} \leftarrow \mathcal{H}(\boldsymbol{\theta}^*, \tilde{\mathbf{u}}, \tilde{\mathbf{x}}_0) + \mathbf{e}, \tag{3.2}$$

where $\tilde{\mathbf{u}} = \{\tilde{u}_1, \ldots, \tilde{u}_n\}$ denotes the forcing data, $\tilde{\mathbf{x}}_0$ signifies the initial states, and $\mathbf{e} = \{e_1, \ldots, e_n\}$ includes observation error (input and calibration data) as well as error due to the fact that the simulator, $\mathcal{H}(\boldsymbol{\theta}^*|\cdot)$ may be systematically different from reality, $\Im(\boldsymbol{\theta}^*)$ for the parameters $\boldsymbol{\theta}^*$. The latter may arise from, e.g., numerical error, spatial discretization and improper (conceptual) model formulation.

Figure 3.1 provides an overview of possible error sources that affect our ability to correctly describe the physical system, $\Im(\boldsymbol{\theta}^*)$ of interest. Forcing data, model parameter, model state, and calibration data error are represented with a probability density function (pdf), whose statistical properties are typically unknown. Errors in the modeled (5) output, $y_{t(t>0)}$ and (6) state, $\mathbf{x}_{t(t>0)}$ dynamics originate from a wide variety of different error sources, including (1) inadequate and/or incomplete knowledge of the model parameters, $\boldsymbol{\theta}^*$ (2) errors in the

input (forcing) data, $\mathbf{u}$ and (3) initial states, $\mathbf{x}_0$ (4) structural inadequacies in the model equations, and/or improper dimensionality of the state space, and (7) errors in the calibration data, $\tilde{y}_{t(t>0)}$. The mathematical operator $\bigotimes$ (also called "likelihood function") is used to judge the distance between the model predictions and corresponding calibration data. This function should explicitly recognize the contribution and role of each individual error source in determining the error residual, but is very difficult to specify correctly with very weak prior information, and hence the pitchfork symbol is used.

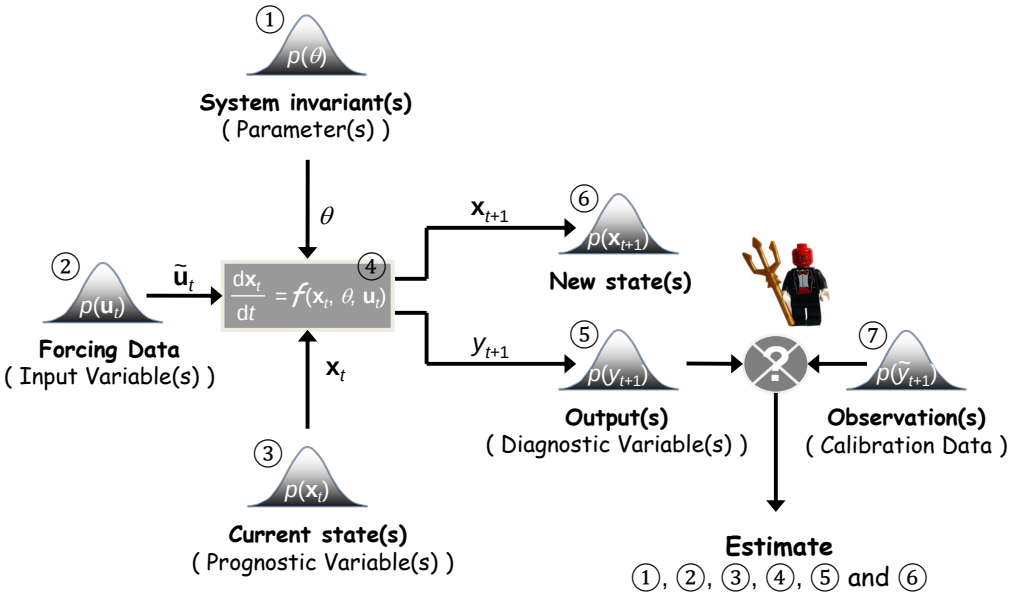


Figure 3.1: Explicit recognition of the role of 1: parameter, 2: forcing data, 3: initial state, 4: model structural, 5: output, 6: state, and 7: calibration data uncertainty. The pitchfork symbol illuminates the difficulty which formulation of the likelihood function (and prior distribution / parameterization of individual error sources) to use to summarize the error residuals. Explicit treatment of individual error sources is required to increase the prospects of explaining the reasons for model inadequacy and learning from the experimental data.

Within the context of hydrologic modeling, measured rainfall depths, and estimates of (potential) evapotranspiration typically constitute the main forcing data. These two input variables strongly determine the simulated streamflow at interior points and the catchment

outlet, surface runoff, soil moisture fluxes and storage of water in the catchment. Examples of model states are soil moisture, groundwater table depth, and hydraulic heads (amongst others). Their knowledge is beneficial to adequately represent the storage of water in the variably saturated zone and groundwater, and hence ensure an adequate model calibration. Finally, calibration data often involves time series of (spatially-distributed) streamflow observations, or time-lapse measurements of tracer concentrations. Inevitably, each of these data sources is subject to uncertainty, which severely complicates parameter estimation and quantification of model structural errors.

During the past 4 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.

Research into error residual distributions had led to the development of a suite of different (hierarchical) likelihood functions for measuring the "closeness" between the model simulations (predictions) and the corresponding data [*Ibbitt and O'Donnell*, 1974, *Sorooshian and Dracup*, 1980, *Kuczera*, 1983a, *Bates and Campbell*, 2001, *Kavetski et al.*, 2006a, *Marshall et al.*, 2007, *Schoups and Vrugt*, 2010a, *Smith et al.*, 2010]. Recent work by *Schoups and Vrugt* [2010a] has resulted in a generalized likelihood function that encapsulates many of the existing likelihood functions in the hydrologic literature, but with additional flexibility

to simultaneously account for correlated, heteroscedastic, and nontraditional error residual distributions.

Research into optimization methods has led to the development of a wide variety of different search methods. Whereas initial approaches utilized local search principles that seek iterative improvement of the objective function from a single starting point in the parameter space [*Ibbitt*, 1972, *Johnston and Pilgrim*, 1976, *Sorooshian and Dracup*, 1980, *Restrepo*, 1982, *Kuczera*, 1983a,b, *Gupta and Sorooshian*, 1983, *Sorooshian et al.*, 1983b, *Troutman*, 1985a,b], problems with parameter insensitivity, curved ridges, local minima, and multiple different regions of attraction has stimulated the development of population based search algorithms that use multiple different points concurrently to locate the global optimum [*Wang*, 1991, *Duan et al.*, 1992, *Yapo et al.*, 1998, *Seibert*, 2000, *Khu and Madsen*, 2005, *Chu et al.*, 2010]. In this regard, the Shuffled Complex Evolution global optimization algorithm of *Duan et al.* [1992] has shown to be effective and efficient in calibrating conceptual watershed models. Recent developments include simple randomized adaptation [*Mazi et al.*, 2004, *Tolson and Shoemaker*, 2007], multimethod ensemble [*Vrugt and Robinson*, 2007a, *Vrugt et al.*, 2009b], and filtering based [*Pauwels*, 2008] parameter estimation methods that further improve search efficiency and reliability.

Research into the information content of data has led to the understanding that it is not the length of the data that matters, but the variability of the observed discharge data [*Kuczera*, 1982, *Sorooshian et al.*, 1983a, *Gupta and Sorooshian*, 1985, *Yapo et al.*, 1996]. Wet and dry periods are both required to make sure that all the different components of the watershed model are excited and the different parameters can be estimated from the calibration data. Post-audit simulations presented in *Vrugt et al.* [2002] using a Bayesian analysis, adaptive Random Walk Metropolis resampling, and value of information (VOI) framework has demonstrated that only a few (daily) streamflow data measurements are necessary to reliably calibrate a conceptual hydrologic model. The remaining data contain

redundant information and could be used to evaluate the reliability of the actual model structure.

Research into numerical solvers has demonstrated that explicit (Euler based) time-stepping schemes introduce considerable streamflow simulation errors and spurious local minima, "pits" and irregularities in the objective function space [*Kavetski et al.*, 2003, 2006c, *Kavetski and Clark*, 2010, *Schoups et al.*, 2010b]. These findings provide a deeper understanding of the convergence problems of local search methods, and demonstrate a need for implicit solvers that iteratively adjust the integration time step based on the state dynamics.

Research into the characterization of uncertainty has resulted in formal and informal statistical approaches. While initial attempts have focused primarily on methods to quantify parameter uncertainty [*Beven and Binley*, 1992, *Freer et al.*, 1996, *Gupta et al.*, 1998, *Kuczera and Parent*, 1998, *Vrugt et al.*, 2002, *Wagener et al.*, 2003, *Beven*, 2006, *Vrugt and Robinson*, 2007a], emerging approaches include state-space filtering [*Vrugt et al.*, 2005, *Moradkhani et al.*, 2005a,b, *Slater and Clark*, 2006, *Reichle*, 2008, *Salamon and Feyen*, 2009, *DeChant and Moradkhani*, 2012, *Vrugt et al.*, 2012], multimodel averaging [*Butts et al.*, 2004, *Georgakakos et al.*, 2004, *Ajami et al.*, 2007], and various (non)Bayesian approaches to treat individual error sources and assess predictive uncertainty [*Montanari and Brath*, 2004, *Vrugt et al.*, 2005, *Kavetski et al.*, 2006a,b, *Kuczera et al.*, 2006, *Huard and Mailhot*, 2006, *Jacquin and Shamseldin*, 2007, *Fenicia et al.*, 2007, *Marshall et al.*, 2007, *Montanari and Grossi*, 2008, *Vrugt et al.*, 2008a,b, *Reichert and Mieleitner*, 2009, *Solomatine and Shrestha*, 2009, *Kuczera et al.*, 2010, *Renard et al.*, 2011a, *Rings et al.*, 2012]. Much progress has also been made in the treatment of forcing data error [*Clark and Slater*, 2006, *Kavetski et al.*, 2006a,b, *Vrugt et al.*, 2008a], development of a formal hierarchical framework to formulate, build and test different watershed models [*Clark et al.*, 2008], and algorithms for efficient sampling of parameter and predictive uncertainty distributions [*Kuczera and Parent*, 1998, *Vrugt et al.*, 2008a, *Kuczera et al.*, 2010, *Laloy and Vrugt*, 2012].

Finally, research into structural adequacy has resulted in data-based mechanistic [*Young*, 2002, 2013], data assimilation [*Vrugt et al.*, 2005, *Smith et al.*, 2008, *Bulygina and Gupta*, 2011], and other stochastic techniques [*Reichert and Mieleitner*, 2009] for inference and iterative refinement of the mathematical structure of conceptual hydrologic models. This has led to the understanding that discharge data contain sufficient information to warrant the identification of a suitable model structure that mimics as closely and consistently as possible the observed watershed behavior at the temporal and spatial scale of measurement.

Most of these development assume input data and model structural errors to be "negligibly small" or to be somehow "absorbed" into the output error residuals. The residuals are then expected to behave statistically similar as the calibration data measurement error. These assumptions are statistically convenient but typically not borne out of the actual probabilistic properties of the residual errors which may show changing bias, variance (heteroscedasticity), skewness, and correlation structures under different hydrologic conditions (and for different parameter sets). 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 likelihood function. For linear systems it is known that ignoring such errors will lead to bias in the estimates of parameter values. The strong and generally difficult to justify assumptions about the nature of the errors have led Beven and coworkers to advocate informal statistical approaches using the Generalized Likelihood Uncertainty Estimation (GLUE) methodology [*Beven and Binley*, 1992, *Beven*, 1993, *Beven and Freer*, 2001, *Beven*, 2006, *Beven et al.*, 2008, *Beven*, 2009].

The origins of the GLUE method lie in trying to deal with uncertainty estimation problems for which simple explicit (theoretical) likelihood assumptions do not seem appropriate. The GLUE methodology rejects the traditional statistical basis for the likelihood function in favor of finding a set of representations (model inputs, model structures, model parameter sets,

model errors) that are behavioral in the sense of being acceptably consistent with the (non-error-free) observations. An informal likelihood measure is used to avoid over conditioning and exclude parts of the model (parameter) space that might provide acceptable fits to the data and be useful in prediction. Since its introduction in 1992, GLUE has found widespread application for uncertainty assessment in many fields of study, including modeling of the rainfall-runoff transformation [*Beven and Binley*, 1992, *Freer et al.*, 1996, *Lamb et al.*, 1998], soil erosion [*Brazier et al.*, 2001], tracer dispersion in a river reach [*Hankin et al.*, 2001], groundwater and well capture zone delineation [*Feyen et al.*, 2001, *Jensen*, 2003], unsaturated zone [*Mertens et al.*, 2004], flood inundation [*Romanowicz et al.*, 1996, *Aronica et al.*, 2002], land-surface-atmosphere interactions [*Franks et al.*, 1997], soil freezing and thawing [*Hanson and Lundin*, 2006], crop yields and soil organic carbon [*Wang et al.*, 2005], and ground radar-rainfall estimation [*Tadesse and Anagnostou*, 2005]. Applications of GLUE are also found in distributed hydrologic modeling [*McMichael et al.*, 2006, *Muleta and Nicklow*, 2005].

In recent years, a strong debate has emerged in the hydrologic community between proponents that adhere strongly to the underlying philosophy of GLUE and believe that the method is a useful working methodology for assessing parameter and predictive uncertainty in non-ideal cases, and researchers and practitioners that strongly oppose incorrect usage of statistics in favor of coherent probabilistic approaches [*Gupta et al.*, 1998, *Beven and Young*, 2003, *Gupta et al.*, 2003, *Christensen*, 2004, *Montanari*, 2005, *Mantovan and Todini*, 2006, *Stedinger et al.*, 2008, *Beven et al.*, 2008, *Beven*, 2009, *Vrugt et al.*, 2008b,c]. This chapter is a follow up of our earlier work [*Vrugt et al.*, 2008c] and demonstrates the similarity of likelihood-free inference used in population and evolutionary genetics [*Pritchard et al.*, 1999, *Beaumont et al.*, 2002], and informal statistical approaches such as GLUE. Likelihood-free inference was introduced in the statistical literature about three decades ago [*Diggle and Gratton*, 1984] for cases where the likelihood is intractable, too expensive to be evaluated, or an explicit formulation is not available. This method is also referred to as Approximate Bayesian Computation (ABC) [*Marjoram et al.*, 2003, *Sisson et al.*, 2007, *Del Moral et al.*,

2008, *Joyce and Marjoram*, 2008, *Grelaud et et al.*, 2009, *Ratmann et al.*, 2009] and widens the class of models for which statistical inference can be performed. This chapter follows a different line of reasoning and approach than *Nott et al.* [2012] who demonstrated that GLUE corresponds to a certain approximate Bayesian procedure even when the "generalized likelihood" is not a true likelihood.

The remainder of this chapter is organized as follows. Section 3.2 briefly summarizes the Bayesian approach to model parameter and predictive uncertainty estimation, with particular emphasis on the choice of the likelihood function used to summarize the probabilistic properties of the error residuals. In Section 3.3 we subsequently introduce likelihood-free computation, and demonstrate the main elements of the ABC procedure by application to a simple Nash-Cascade series of three linear reservoirs. This is followed by Section 3.4 in which the conceptual and statistical equivalence of ABC and the limits of acceptability approach of GLUE is demonstrated. Section 3.5 then proceeds with a comparison between GLUE and ABC using the Sacramento soil moisture accounting (SAC-SMA) model [*Burnash et al.*, 1973] and hmodel [*Schoups and Vrugt*, 2010a] and discharge data form two contrasting watersheds in the US. Finally, in Section 3.6 we summarize our results and discuss the main findings.

3.2 Bayesian Inference

Bayes theorem is a simple rule how to update the prior probability of a certain hypothesis when new, relevant information (data) is becoming available. The term "Bayesian" refers to the 18th century mathematician and Reverend Thomas Bayes (1701-1761), who studied how to compute a distribution for the probability parameter of a binomial distribution. If we conveniently assume that the model parameters are the only source of uncertainty, we

can write Bayes' rule as follows

$$p(\boldsymbol{\theta}|\tilde{\mathbf{Y}}) = \frac{p(\boldsymbol{\theta})p(\tilde{\mathbf{Y}}|\boldsymbol{\theta})}{p(\tilde{\mathbf{Y}})}, \tag{3.3}$$

where $p(\boldsymbol{\theta})$ ($p(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$) signifies the prior (posterior) parameter distribution, $\mathcal{L}(\boldsymbol{\theta}|\tilde{\mathbf{Y}}) \equiv p(\tilde{\mathbf{Y}}|\boldsymbol{\theta})$ denotes the likelihood function, and $p(\tilde{\mathbf{Y}})$ represents the evidence (or normalization constant).

A key task is now to summarize the posterior distribution, $p(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$ for example by the mean, the covariance or percentiles of individual model parameters. When this task cannot be carried out by analytical means or analytical approximation, iterative methods are needed to generate a sample from the posterior distribution. The desired summary is then obtained from this sample. Knowledge of the normalization constant, $p(\tilde{\mathbf{Y}})$ is not required for sampling of the parameters as all our statistical inferences can be made from the unnormalized density. Explicit knowledge of $p(\tilde{\mathbf{Y}})$ is however desired for Bayesian model selection and averaging.

Epistemic uncertainty (model inadequacy due to a lack of knowledge) and forcing data errors can, in principle, be summarized using hyper parameters (latent variables) and their distribution derived from Bayesian inference. Yet, practical experience suggests that it is very difficult, if not impossible, to disentangle the effects of individual error sources particularly if a single data type, $\tilde{\mathbf{Y}}$ constitutes the only basis for model evaluation. In our previous chapter (Chapter 2) we have demonstrated the ability of likelihood-free inference methods to significantly enhance the chances of detecting model structural deficiencies. This approach marks a progress towards improving our perceptual-conceptual-theoretical view(s) of the world, expressed as model structural hypotheses (assumptions and conjectures).

The scope of the present chapter is fundamentally different than that of our previous work published in *Vrugt and Sadegh* [2013]. We show herein, that ABC has many elements in common with the limits of acceptability approach of GLUE, but benefits from a much better

statistical underpinning. To be consistent with GLUE, we map the input-output uncertainty on the model parameters. Under ideal conditions with an adequate model and perfect forcing data, the error residuals follow a zero-mean Gaussian distribution

$$\mathcal{L}(\boldsymbol{\theta}|\tilde{\mathbf{Y}}, \tilde{\mathbf{u}}, \tilde{\mathbf{x}}_0) = \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\hat{\sigma}^2_{\tilde{\mathbf{Y}}}}} \exp\left[-\frac{1}{2}\hat{\sigma}^{-2}_{\tilde{\mathbf{Y}}}\big(\tilde{y}_t - y_t(\boldsymbol{\theta}, \tilde{\mathbf{u}}, \tilde{\mathbf{x}}_0)\big)^2\right], \tag{3.4}$$

and $\boldsymbol{\theta}$ should converge to $\boldsymbol{\theta}^*$ where $\hat{\sigma}_{\tilde{\mathbf{Y}}}$ is an estimate of the standard deviation of the measurement error. The value of $\hat{\sigma}_{\tilde{\mathbf{Y}}}$ can be specified *a-priori* based on knowledge of the measurements errors, or alternatively its value can be inferred simultaneously with the values of $\boldsymbol{\theta}$ [*Vrugt et al.*, 2008b, *Bikowski et al.*, 2012, *Laloy and Vrugt*, 2012]. It is worth noting that the data often come from only a single experiment. So while it is possible to quantify numerical errors, such as those due to discretization (see *Kaipio et al.* [2004], *Nissinen et al.* [2009]), there is no opportunity to control the boundary conditions of (large-scale) natural systems to obtain data from additional experiments in which some controllable inputs have been varied.

The likelihood function, $\mathcal{L}(\cdot)$ in Equation (3.4) is useful for simple regression problems, but the assumption of independent identically distributed Gaussian error residuals cannot be justified in environmental modeling. The presence of input data and model structural errors introduces complex error residual distributions whose probabilistic properties are difficult to describe accurately with classical likelihood functions. The choice of an adequate likelihood function, $L(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$ has therefore been the subject of considerable debate in the hydrologic and statistical literature. In response to this, *Schoups and Vrugt* [2010a] has introduced a generalized likelihood function that better extends the applicability of commonly likelihood functions to situations where residual errors are correlated, heteroscedastic, and non-Gaussian with varying degrees of kurtosis and skewness. Application to daily rainfall-runoff data from a dry and humid basin showed that (1) residual errors are much better described by a heteroscedastic, first-order, auto-correlated error model with a Laplacian distribution function

characterized by heavier tails than a Gaussian distribution; and (2) compared to a standard least-squares approach, proper representation of the statistical distribution of residual errors yields tighter predictive uncertainty bands and different parameter uncertainty estimates that are less sensitive to the particular time period used for inference, (3) multiplicative bias factors improve the prediction of peak flow, and (4) near zero-flows are better described with a skewed error distribution.

The generalized likelihood function improves the statistical description of the error residuals, yet it does not separate out the effect of individual error sources. Another, from the viewpoint of this chapter less important deficiency is that the use of a single performance metric, $\mathcal{L}$ no matter how carefully chosen, is inadequate to extract all information from the available calibration data. The use of such "insufficient statistic" promotes equifinality, and makes it unnecessarily difficult to find the preferred parameter values. This is not desirable and explains why calibration of highly-parameterized models is often found to be very time consuming and difficult.

3.3 Approximate Bayesian Computation

Whereas traditional Bayesian approaches require us to specify an explicit likelihood function, $\mathcal{L}(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$, ABC approaches avoid explicit evaluation of the likelihood function in favor of (a set of) summary variables that better extract the information from the available data. The premise behind ABC is that $\boldsymbol{\theta}'$ should be a sample from the posterior distribution as long as the distance between the observed and simulated data, hereafter referred to as $\rho\big(\tilde{\mathbf{Y}}, \mathbf{Y}(\boldsymbol{\theta}')\big)$ is less than some small value, ϵ. For sufficiently complex models and large data sets the probability of happening upon a simulation run that yields precisely the same data set as the one observed will be very small, often unacceptably so. So rather than considering the data, $\tilde{\mathbf{Y}}$ itself we consider a summary statistic of the data, $S(\tilde{\mathbf{Y}})$, and use a distance function

[*Marjoram et al.*, 2003, *Sisson et al.*, 2007]

$$\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}^{'}))\big) \leq \epsilon, \tag{3.5}$$

to decide whether to accept the parameter values, $\boldsymbol{\theta}_i$ or not. A pseudo-code of the generic ABC approach is given below.

Algorithm 4 Rejection sampler (ABC-REJ)

for $i = 1, \ldots, N$ **do**

  repeat

    generate $\boldsymbol{\theta}^{'}$ from the prior distribution, $p(\boldsymbol{\theta})$

    simulate $\mathbf{Y}$ from the model, $\mathbf{Y} \leftarrow \mathcal{H}(\boldsymbol{\theta}^{'}|\cdot)$

  until $\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}))\big) \leq \epsilon$

  set $\boldsymbol{\theta}_i = \boldsymbol{\theta}^{'}$

end for

In words, the ABC algorithm proceeds as follows. First we sample a candidate point, $\boldsymbol{\theta}^{'}$ from some prior distribution, $p(\boldsymbol{\theta})$. We then use this proposal to simulate the output of the model, $\mathbf{Y} \leftarrow \mathcal{H}(\boldsymbol{\theta}^{'}|\cdot)$ and use this n-vector to calculate one or multiple summary metrics. A distance function, $\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}^{'}))\big)$ is then used to decide whether to accept $\boldsymbol{\theta}^{'}$ or not. If this distance function is smaller than some pre-defined tolerance value, ϵ then the simulation is close enough to the observations that the candidate point, $\boldsymbol{\theta}^{'}$ has some nonzero probability of being in the approximate posterior distribution, $\hat{p}\left(\boldsymbol{\theta}|\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y})\big) \leq \epsilon\right)$. This algorithm converges to the true posterior $p(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$ when $\epsilon \to 0$, provided that the summary statistic(s), $S(\cdot)$ is (are) near sufficient [*Pritchard et al.*, 1999, *Beaumont et al.*, 2002, *Ratmann et al.*, 2009, *Turner and van Zandt*, 2012].

The choice of summary metrics is obviously an important consideration in the application of

ABC. These criteria (signatures) should be chosen so that the loss of information from the original data is minimized. Information theory can help to determine such (set of) sufficient statistic(s) [*Barnes et al.*, 2011], but this is beyond the scope of the present study (as will soon become evident) and will therefore be a main focus in future publications.

Another issue that deserves careful attention is that ABC can only be used with a stochastic model operator. Otherwise the posterior parameter distribution will continue to shrink in the limit of ϵ going to zero, and eventually converge to a Dirac delta function (single point) if the model is sufficiently adequate. In theory, we should therefore corrupt the output of the deterministic model, $\mathbf{Y} \leftarrow \mathcal{H}(\boldsymbol{\theta}|\cdot)$ with a random (measurement) error, but this is deemed unnecessary within the present context. We will revisit this important issue in the penultimate paragraph of this chapter. The ABC findings presented in Tables 3.1-3.6 and Figures 3.3-3.14 thus pertain to deterministic model output only.

To illustrate the ABC methodology, we consider a Nash-Cascade instantaneous unit hydrograph. This model routes inflow (rainfall) through a series of linear reservoirs that all have the same recession coefficient. Mathematically, this cascade of m linear reservoirs with recession coefficient r can be written as follows [*Nash*, 1960]

$$h_t(r,m) = \frac{1}{r\Gamma(m)} \left(\frac{t}{r}\right)^{(m-1)} \exp\left(-\frac{t}{r}\right), \tag{3.6}$$

where t (days) denotes time, $\Gamma(\cdot)$ signifies the gamma function, and $h_t(\cdot)$ is the modeled response at time t. A 365-day period with synthetic daily streamflow data (in m^3/s) was generated by driving the Nash-Cascade model of Equation (3.6) with an artificial precipitation record. We assume $m = 3$ reservoirs, and a recession constant of $r = 2$ days. This artificial data set is subsequently perturbed with a heteroscedastic error (non-constant variance) with standard deviation equal to 20% of the original simulated discharge values. Figure 3.2 plots the original simulated discharge time series (blue line) and the corrupted

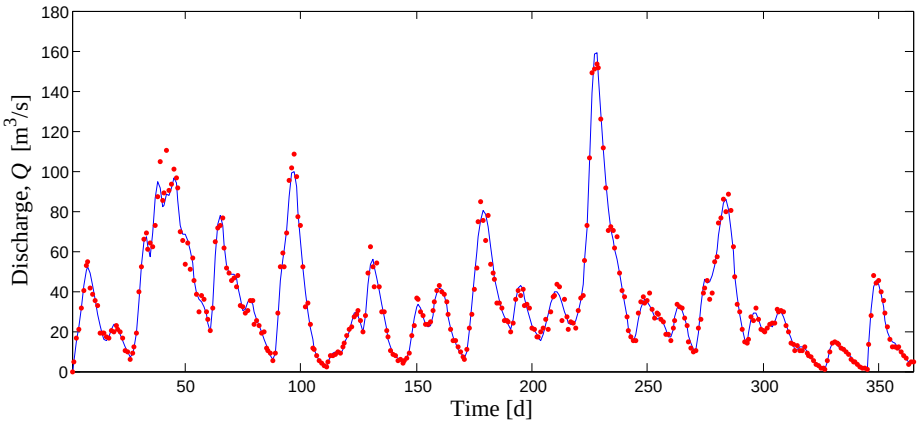


Figure 3.2: Synthetic discharge time series (blue line) simulated with the Nash Cascade model, and the corrupted observations (red points) used in the GLUE and ABC analysis.

observations (red circles) used in the ABC analysis to derive the posterior distribution of the recession constant.

We are now left with a selection of the summary statistic, $S(\cdot)$ to decide whether a candidate point (model simulation) is behavioral or not. For illustrative purposes we start with the mean of the actual data,

$$\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}))\big) = \big|\text{mean}(\tilde{\mathbf{Y}}) - \text{mean}(\mathbf{Y}(\boldsymbol{\theta}))\big| \tag{3.7}$$

to estimate the posterior distribution of the recession constraint. This metric is rather weak and cannot be considered "sufficient", but for now it shows how the ABC methodology works in practice.

A uniform prior with $r \in [0, 4]$ was used in all our calculations. To increase computational efficiency, we used an improved variant of the ABC population Monte Carlo (PMC) scheme of *Turner and van Zandt* [2012], the details of which appear in Appendix A. In short, the PMC sampler starts out as 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 decreased and the proposal distribution, $q_j(\boldsymbol{\theta}_k^{j-1}, \cdot) = N_d(\boldsymbol{\theta}_k^{j-1}, \sum^j)_{(j>1)}$ adapted using $\sum^j = \text{Cov}(\boldsymbol{\theta}_1^{j-1}, \ldots, \boldsymbol{\theta}_N^{j-1})$ with $\boldsymbol{\theta}_k$ drawn from a multinomial distribution, $\mathfrak{F}(\boldsymbol{\theta}_{1:N}^{j-1} | \mathbf{w}_{1:N}^{j-1})$

where $\mathbf{w}_{1:N}^{j-1}$ denote the posterior weights ($w_l^{j-1} \geq 0;\ \sum_{l=1}^{N} w_l^{j-1} = 1$). 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, $p(\boldsymbol{\theta})$ is a poor approximation of the actual posterior distribution.

The PMC sampler of *Turner and van Zandt* [2012] assumes that the sequence of ϵ values is specified by the user. This does not necessarily lead to the most efficient search. Our sampler therefore adaptively determines the next value of $\epsilon_j; j > 1$ from the cumulative distribution function of the $\rho(\cdot)$ values of the N most recent accepted samples. Details of this procedure are given in Appendix B. For the present case study, an initial value of $\epsilon = 1$ is used, and this value is adaptively decreased until a value of $\epsilon = 0.05$ is reached. Lower values of ϵ provide similar posterior estimates, yet unnecessarily increase the computational burden of the ABC analysis [*Vrugt and Sadegh*, 2013].

Figure 3.3A presents a histogram of the posterior marginal distribution of r derived from the ABC-PMC analysis using the mean observed flow as summary statistic. The red square denotes the true parameter value used to create the synthetic data. For completeness we also present in the middle panel the results of DREAM [*Vrugt et al.*, 2008a, 2009a] using the likelihood function of Equation (3.4) with a heteroscedastic measurement error, $\hat{\boldsymbol{\sigma}}_{\tilde{\mathbf{Y}}} = 0.2\tilde{\mathbf{Y}}$.

Perhaps not surprisingly, the ABC-derived posterior distribution is poorly defined by calibration against the mean observed discharge value. The behavioral recession constants extend a larger portion of its prior distribution. This suggests that the observed (synthetic) discharge data do not contain information about the recession constant of the three reservoirs. This finding is perhaps not surprising. Many different values of the recession constant, r can be found with mean simulated discharge value similar to that of the observed data, but

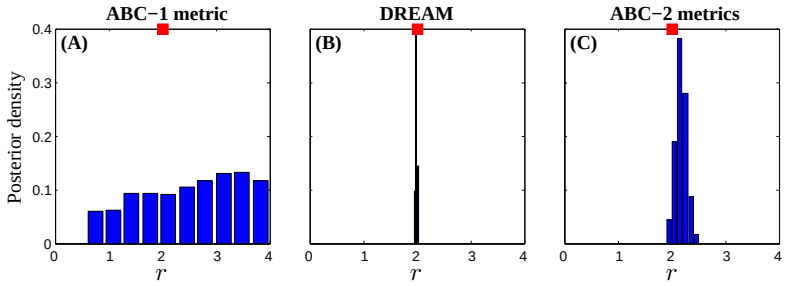


Figure 3.3: Posterior distribution of the recession parameter derived from (A) ABC using the mean of the data as summary metric, (B) DREAM using a heteroscedastic measurement error, and (C) ABC using the mean and standard deviation of the data as summary statistics. The true parameter value is indicated with a red cube on top of each plot.

with poor accuracy of the simulated streamflow dynamics. Indeed, if a classical likelihood function is used (Figure 3.3B) the recession constant is much better defined with maximum a-posteriori density equal to $r = 2$, and 95% posterior parameter uncertainty ranges that vary between 1.95 and 2.05.

Fortunately, nothing prevents us from using more than one summary statistic in the ABC-analysis to measure different and complementary parts of model behavior. To be meaningful in practice, such statistics should preferably measure hydrologically relevant signatures of watershed behavior. Such approach was introduced in our previous work [*Vrugt and Sadegh*, 2013], and for simplicity we now augment the first metric (mean of the data) with another simple statistic (standard deviation of data)

$$\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}))\big) = \max\left(\Big|\text{mean}(\tilde{\mathbf{Y}}) - \text{mean}(\mathbf{Y}(\boldsymbol{\theta}))\Big|, \Big|\text{std}(\tilde{\mathbf{Y}}) - \text{std}\left(\mathbf{Y}\left(\boldsymbol{\theta}\right)\right)\Big|\right), \quad (3.8)$$

to decide whether the model simulation can be considered behavioral or not. The results of this analysis are shown in Figure 3.3C using a minimum value of $\epsilon = 0.05$. The recession constant appears much better defined, but the width of the (marginal) posterior distribution is still considerably larger than what can be expected from a classical likelihood function using MCMC simulation with DREAM (Figure 3.3B). This simply conveys that our two summary metrics are jointly insufficient, and that, if so desired, more powerful metrics

should be used.

The ABC methodology allows the use of a wide arsenal of summary metrics and distance functions to judge the distance between the model simulation and observations. Common examples in genetics include the Canberra, Euclidean, and Manhattan distance. Those are readily applied in hydrology as well, including summary statistics such as the Nash-Sutcliffe efficiency (NSE) [*Nash and Sutcliffe*, 1970], Mean Square Error (MSE), and others listed in Table 1 of *Gupta et al.* [1998]. Temporal disaggregation of the data and model simulations would preserve the statistical moments of $\tilde{\mathbf{Y}}$ such as the mean, median, standard deviation, kurtosis, and skewness. The use of flow duration curves could be beneficial in this regard as characteristic of the watersheds response to rainfall [*Vrugt and Sadegh*, 2013].

3.4 Statistical Equivalence of ABC and GLUE

Now the basic principles of ABC have been discussed in some detail using the simple one-parameter unit hydrograph toy problem, it is not difficult to see the many similarities of GLUE and ABC

1. The distance function specified in Equation (3.5) has many elements in common with the informal threshold used in the classical GLUE approach to differentiate between behavioral and nonbehavioral samples. This is perhaps more obvious if we use the following notation

$$\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}))\big) = \sum_{t=1}^{n} \mathrm{I}\big(|\tilde{y}_t - y_t(\boldsymbol{\theta})| \leq \delta_t\big), \tag{3.9}$$

where $\mathrm{I}(A)$ is a simple indicator function that is "1" if A is true, and "0" otherwise, and $\delta_t; t = \{1, \ldots, n\}$ constitutes the effective observation error that takes into account multiple sources of error [*Beven*, 2006]. This value is defined a-priori by the

user. The ABC approach can thus be made mathematically equivalent to the limits of acceptability approach of GLUE if each observation is used as summary statistic.

2. The ABC-REJ sampler is similar to the Latin Hypercube sampling strategy used in GLUE to find behavioral solutions. Both methods use a fixed proposal distribution to sample from the prior parameter distribution. If the corresponding simulation falls within the bounds specified by the effective observation error, then the parameter combination will be classified as behavioral, otherwise, the proposal will be rejected. Sampling continues until N behavioral solutions are found.

Following the first proposition, a solution is deemed behavioral with ABC if its simulated discharge time series falls within the interval, $[\tilde{y}_t - \delta_t, \tilde{y}_t + \delta_t]$ for $t = \{1, \ldots, n\}$. This is similar to the limits of acceptability approach of GLUE if a simple discrete (0/1) membership function is used. For the synthetic toy example used herein, we define the effective observation error to be $\boldsymbol{\delta} = \alpha \times \hat{\boldsymbol{\sigma}}_{\tilde{y}}$ with $\alpha = 2$. This is equivalent to $\delta_t = 0.4\tilde{y}_t$.

The goal of the ABC analysis now becomes finding all those parameter combinations that consistently fall within the effective observation error of the discharge data, and hence receive a perfect score of Equation (3.9) equal to n. This constitutes a maximization problem, and differs from a typical implementation of ABC where the distance to the summary statistics, and value of ϵ is being minimized. In our numerical implementation with the PMC sampler we therefore adapt Equation (3.9) and calculate

$$\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}))\big) = \frac{1}{n}\left(n - \sum_{t=1}^{n} \mathrm{I}\big(|\tilde{y}_t - y_t(\boldsymbol{\theta})| \leq \delta_t\big)\right), \tag{3.10}$$

to decide whether a simulation is behavioral or not. For a perfect simulation, $\rho\big(\cdot\big)$ will be zero. But in most practical applications it is not possible to find a simulation that satisfies $\epsilon = 0$. For instance, for the present Nash-Cascade toy example with $\alpha = 2$ and thus $\delta_t = 0.4\tilde{y}_t;\ t = \{1, \ldots, n\}$, a minimum value of $\rho\big(\cdot\big)$ of about 0.05 is to be expected. This

follows directly from statistical theory (about 95% of the observations are included in the interval of 2 times the standard deviation).

The adaptive updating strategy of ϵ in PMC not only guarantees a more efficient search strategy than ABC-REJ (GLUE), but also automatically determines the maximum attainable coverage of the discharge observations within the limits of acceptability. In the first iteration, we set $\epsilon = 0.75$ and thus define a behavioral solution as one that contains at least 25% of the observed discharge data within the interval, $[y_t - \delta_t, y_t + \delta_t]_{(t=\{1,\ldots,n\})}$. During each successive next iteration the value of ϵ is sequentially reduced, and the PMC sampler terminates when the difference between two subsequent ϵ values is less than 0.02, or in mathematical notation $\epsilon_j - \epsilon_{j-1} < 0.02$. In all our simulations presented herein we request PMC to create $N = 1,000$ behavioral solutions at each different ϵ value (iteration). We report our results for $\epsilon \leq 0.10$.

Figure 3.4A presents the results of the ABC-PMC analysis and plots the 95% streamflow simulation uncertainty ranges (dark grey region) using the ABC-PMC sampler. This result is derived from the $N = 1,000$ posterior solutions using the 2.5 and 97.5 percentile of the simulated discharge values. The artificial discharge observations are indicated with red circles. The simulations nicely track the observed data with uncertainty intervals that appear relatively narrow and encompass about 90% of the data. The upper panel plots the results for GLUE using the same limits of acceptability. Latin hypercube sampling was used to create $1,000$ behavioral solutions at $\epsilon = 0.10$ using an algorithm virtually identical to that of ABC-REJ. Perhaps not surprising, the results are identical to those presented for ABC. Although the numerical results are identical, the computational efficiency of both methods is not. The ABC-PMC sampler exhibits an acceptance rate of about 53.5% whereas for GLUE (and hence ABC-REJ) this is about 17.0%.

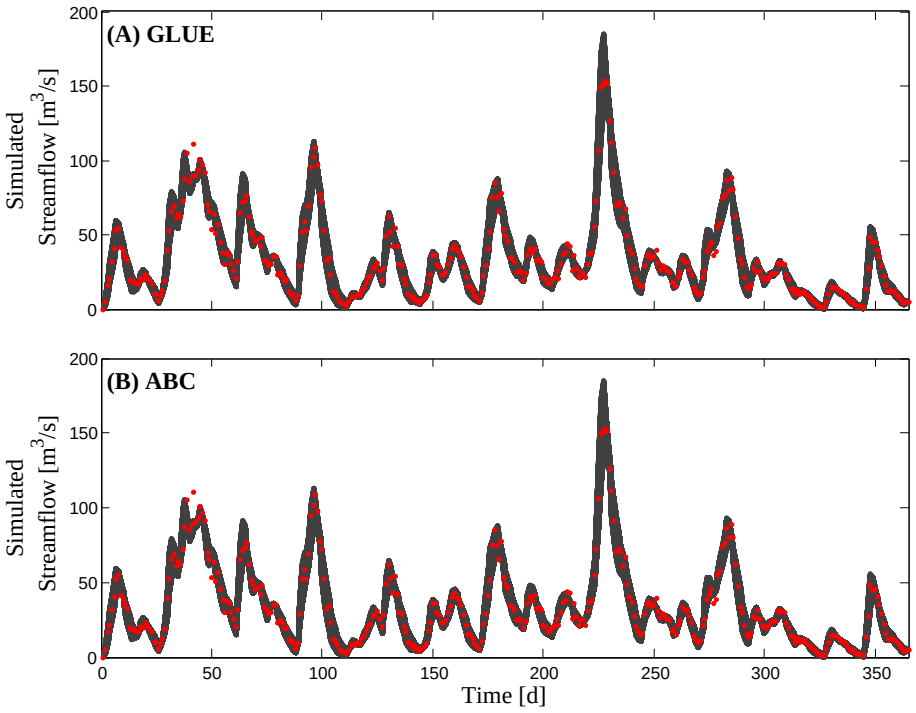


Figure 3.4: 95% streamflow uncertainty ranges (dark region) derived from GLUE (top panel) and ABC (bottom panel). The red points mark the actual discharge observations. The prediction uncertainty ranges derived with both methods appear virtually identical and nicely capture the desired percentage of streamflow observations.

3.5 Case Studies: Hydrologic Modeling

Now the ABC method has been discussed in some detail and the theoretical connection of this approach with GLUE has been demonstrated, we proceed with numerical simulation using five years of daily streamflow data from the French Broad river basin at Ashvile, North Carolina (Jan. 1, 1962 to Dec. 30, 1966) and the Leaf River (Oct. 1, 1952 to Sep. 30, 1957) north of Collins, Mississippi. These watersheds have been studied extensively in the literature and details of the data can be found in related publications. Two lumped conceptual hydrologic models are used to describe the rainfall-runoff transformation. This includes the 7-parameter hmodel described in detail in *Schoups and Vrugt* [2010a] and the 13-parameter SAC-SMA model [*Burnash et al.*, 1973]. Inputs to these models include mean areal precipitation (MAP) and potential evapotranspiration (PET) while the outputs are

Table 3.1: Prior uncertainty ranges of hmodel parameters.

Parameter	Symbol	Minimum	Maximum	Units
Maximum 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

estimated evapotranspiration and channel inflow. Numerical, conceptual, and computational details of both models can be found in the cited publications, and so will not be repeated herein. Tables 3.1 and 3.2 summarize the parameters of both models and their upper and lower bound values.

Implementation of the limits of acceptability approach requires knowledge of the effective observation error. This error varies dynamically with flow level and constitutes the combined effect of input data, model structural and calibration data measurement error. In practice, the user defines the limits of acceptability for each individual observation, but for convenience we follow a different approach and set the effective observation error as a multiple of the actual discharge measurement error. We follow *Vrugt et al.* [2005] and use consecutive differences of the calibration observations to calculate the measurement data error (e.g. 7 in Figure 3.1)

$$\hat{\sigma}_{\tilde{\mathbf{Y}}} = \sqrt{\binom{2l}{l}^{-1} (\Delta^l \tilde{y}_t)^2}, \tag{3.11}$$

where Δ^l denotes the difference operator applied l subsequent times [*Rice*, 1984, *Hall et al.*, 1990, *Seifert et al.*, 1993, *Dette et al.*, 1998]. This estimator was introduced in *Vrugt et al.* [2005] and shown to work well for daily and hourly discharge data. Heteroscedasticity is easily identified by applying the nonparametric estimator locally in the calibration data time series.

This provides a n-vector of measurement errors, hereafter referred to as $\hat{\boldsymbol{\sigma}}_{\tilde{\mathbf{Y}}} = \{\hat{\sigma}_{\tilde{y}_1}, \ldots, \hat{\sigma}_{\tilde{y}_n}\}$. The limits of acceptability in Equation (3.9) are now defined to be, $\delta_t = \alpha \times \hat{\sigma}_{\tilde{y}_t}$ for $t = \{1, \ldots, n\}$ using $\alpha = 2$.

Table 3.2: Description of the SAC-SMA model parameters and their (uniform) prior uncertainty ranges.

Parameter	Symbol	Minimum	Maximum	Units
Upper zone tension water maximum storage	UZTWM	1.0	150.0	mm
Upper zone free water maximum storage	UZFWM	1.0	150.0	mm
Lower zone tension water maximum storage	LZTWM	1.0	500.0	mm
Lower zone free water primary maximum storage	LZFPM	1.0	1000.0	mm
Lower zone free water supplemental maximum storage	LZFSM	1.0	1000.0	mm
Additional impervious area	ADIMP	0.0	0.40	-
Upper zone free water lateral depletion rate	UZK	0.1	0.5	day^{-1}
Lower zone primary free water depletion rate	LZPK	0.0001	0.025	day^{-1}
Lower zone supplemental free water depletion rate	LZSK	0.01	0.25	day^{-1}
Maximum percolation rate	ZPERC	1.0	250.0	-
Exponent of the percolation equation	REXP	1.0	5.0	-
Impervious fraction of the watershed area	PCTIM	0.0	0.1	-
Fraction percolating from upper to lower zone free water storage	PFREE	0.0	0.6	-

We now summarize the results of GLUE and ABC for both models and watersheds. Figure 3.5 plots histograms of the behavioral solutions of an illustrative set of five SAC-SMA model parameters for the French Broad watershed. The PMC sampler terminated its search with $\epsilon \leq 0.06$ corresponding to a coverage of 94% of the discharge data within the effective observation error. The top panel presents the results for GLUE (limits of acceptability) and the bottom panel shows the corresponding counterparts for ABC. To limit the computational burden, GLUE was terminated after 300 behavioral solutions were found. This is sufficient for comparative purposes. The marginal distribution of the lower zone primary free water depletion rate (LZPK) follows a normal distribution, whereas the histograms of the other parameters deviate considerably from normality and tend to assign the highest probability mass at the lower (PCTIM, ADIMP and LZFSM) or upper bound (LZFPM). The posterior parameter uncertainty ranges appear rather large, and essentially cover the entire prior distribution defined previously in Table 3.2. This uncertainty is perhaps unrealistically large and much larger than what can be expected from an explicit likelihood function, but not surprising given the size of the effective observation error used to define the limits of acceptability. What is most important however is the finding that the GLUE and ABC derived posterior parameter distributions are essentially similar. This provides further support for our claim that the limits of acceptability approach of GLUE can be interpreted as a special case of formal Bayes. We will further elaborate on this equivalence in the final section of this chapter.

Now the posterior parameter uncertainty has been defined, we focus our attention on the actual discharge simulations. Figure 3.6 presents the outcome of this analysis, and presents the 95% streamflow uncertainty ranges (gray region) of the GLUE (top panel) and ABC (bottom panel) derived posterior parameter distribution. The simulation uncertainty ranges appear rather large, but nicely cover approximately 74% of the discharge observations. The simulation results of GLUE and ABC are in strong agreement, which is to be expected given the strong similarity of the behavioral samples derived with both methods.

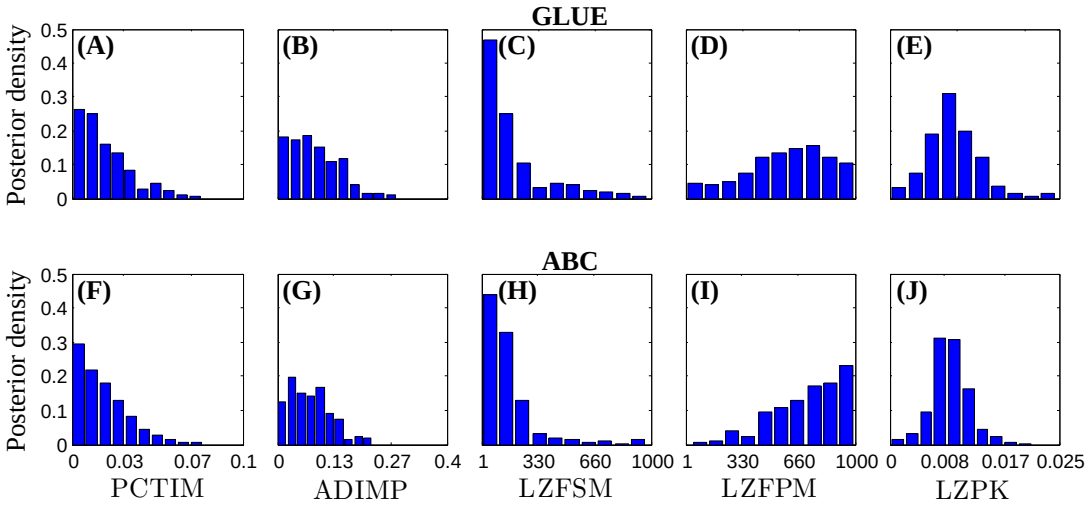


Figure 3.5: Posterior distribution of five randomly chosen SAC-SMA model parameters derived from (A) GLUE, and (B) ABC using historical streamflow data from the French Broad river basin.

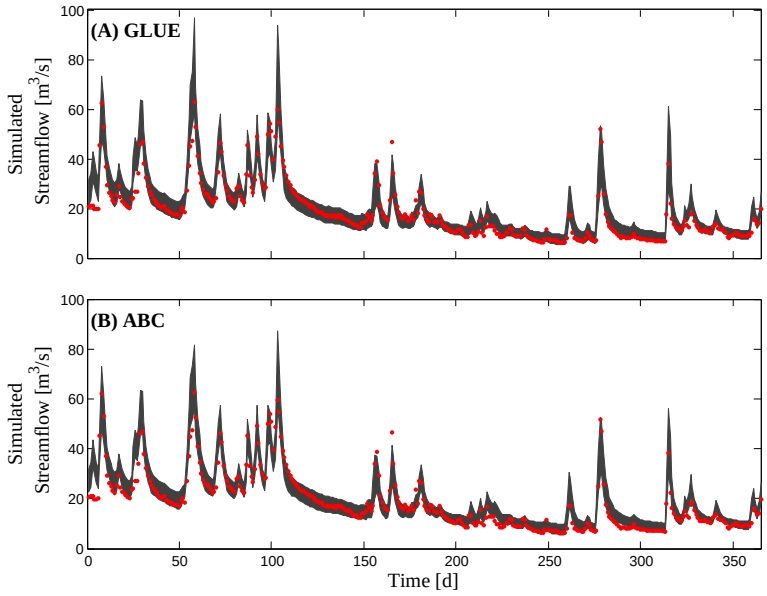


Figure 3.6: SAC-SMA derived 95% streamflow simulation uncertainty ranges (grey region) of the calibration period of the French Broad river basin using (A) GLUE and (B) ABC. The observed discharge data are indicated with the solid red dots. We limit our display to the first 365-days of the calibration data set to simplify graphical interpretation. The simulation uncertainty ranges appear very similar and nicely cover the observed discharge data.

Although the numerical results of GLUE and ABC are very similar, the PMC sampler requires only 1/30 (1/8) of the simulations of GLUE to locate $N = 1,000$ posterior solutions for the SAC-SMA (hmodel) (see Table 3.3). The advantage of PMC is more and more appar-

Table 3.3: Computational efficiency of GLUE and ABC for the French Broad data set: Acceptance rate (AR, %) and total number of SAC-SMA and hmodel function evaluations (FE) required to sample $N = 1,000$ behavioral solutions.

	SAC-SMA		hModel	
	ABC	GLUE	ABC	GLUE
AR, %	0.41	0.016	0.50	0.06
FE, -	242,004	6,110,640**	201,192	1,608,810**

**Derived from linear scaling of FE needed to sample 300 (SAC-SMA) and 100 (hmodel) behavioral solutions.

ent with increasing dimensionality of the parameter space. If the search space is relatively low-dimensional (hmodel) and the space of behavioral solutions relatively large in comparison to the prior parameter space, both sampling methods will rapidly sample $N = 1,000$ behavioral solutions. If, on the contrary, the search space is of higher dimensions (SAC-SMA), or the behavioral solution space is made up of a small portion of the prior parameter space, Latin hypercube sampling (and ABC-REJ) will be rather inefficient needing many thousands of draws from the prior distribution to find just a handful of good (behavioral) solutions. The PMC sampler achieves a higher sampling efficiency by iteratively reducing the value of ϵ during the search, and adaptively updating the scale and orientation of the proposal distribution. Note that the PMC and Latin Hypercube sampling strategies used herein vary all parameters at a time, and hence further efficiency improvements are to be expected in high-dimensional parameter spaces with the user of genetic operators such as crossover and mutation.

To provide more insights into the sampled parameter space of the French Broad river basin, please consider Figure 3.7 that presents two-dimensional scatter plots of the posterior samples derived with GLUE (top panel) and ABC-PMC (bottom panel) for three selected parameter pairs. The bivariate sample plots appear very similar and confirm our previous findings in Figure 3.5 and demonstrate significant scatter with behavioral samples that extend their entire uniform prior ranges. But this does not necessarily mean that the posterior parameter

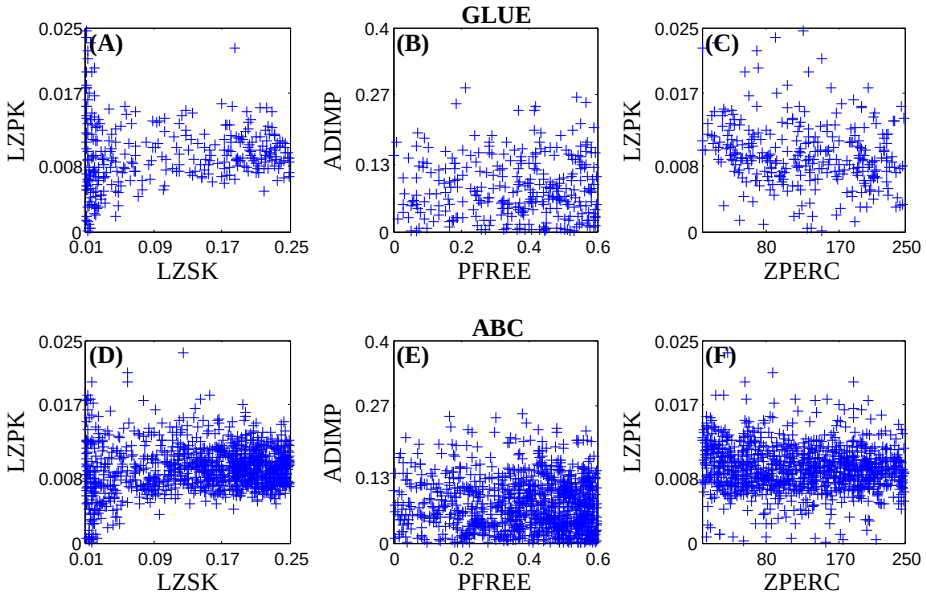


Figure 3.7: Bivariate scatter plots of the behavioral (posterior) samples of three different (randomly selected) parameter pairs using GLUE (top panel) and ABC (bottom panel): (A) LZSK-LZPK, (B) PFREE-ADIMP, and (C) ZPERC-LZPK. The scatter plots derived with both methods are in close agreement but demonstrate an important difference in sampling density. The computational budget of GLUE was limited to approximately 2 days, and within this time frame the Latin hypercube sampling method located only 300 behavioral solutions.

space is badly defined. Instead, large portions of the (A) LZSK-LZPK, (B) PFREE-ADIMP, and (C) ZPERC-LZPK subspaces are virtually empty and thus deemed nonbehavioral. This suggests at least some level of correlation between the posterior parameter samples. The difference in sampling density between both panels is simply due to an insufficient computational budget for GLUE to create $N = 1,000$ behavioral solutions. GLUE was terminated after 300 posterior samples were found.

We now proceed with out-of-sample prediction, and plot in Figure 3.8 the streamflow uncertainty ranges (gray region) of the SAC-SMA model for a three year portion of the evaluation data set of the French Broad watershed. This period commences immediately after the last day of the calibration data set, and the initial state at $t = 0$ have been derived from the calibration ensemble. The top panel presents the results of GLUE whereas the bottom panel plots the corresponding results of ABC. Perhaps not surprisingly, both methods exhibit

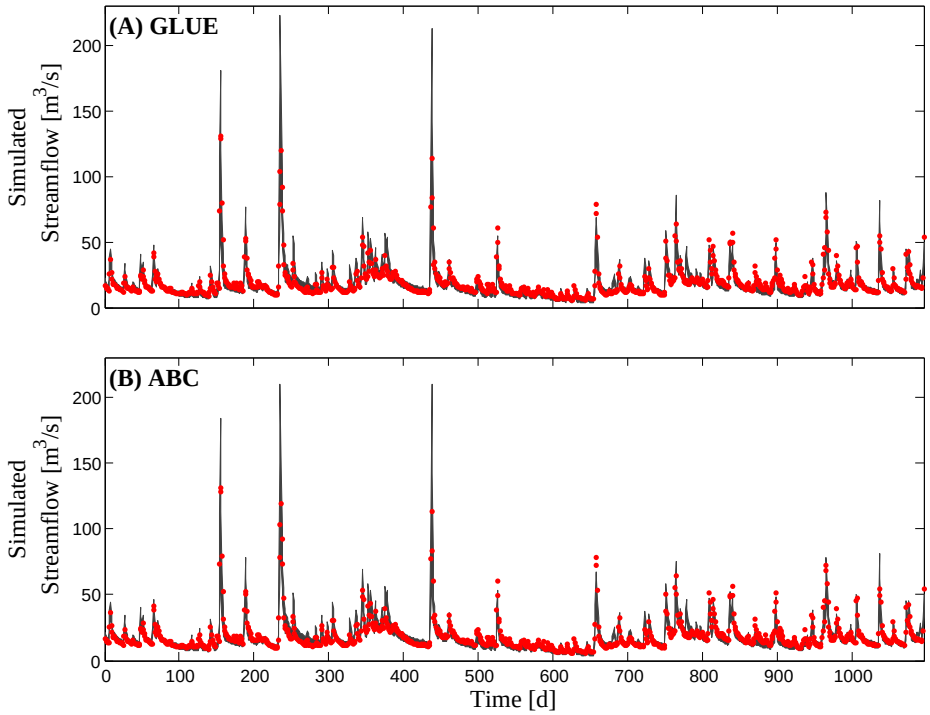


Figure 3.8: SAC-SMA derived 95% streamflow simulation uncertainty ranges (grey region) for a three-year portion of the evaluation period of the French Broad river basin using the (A) GLUE and (B) ABC derived posterior parameter distribution. The observed discharge data are indicated with the solid red dots.

similar results and provide a discharge ensemble that envelops about 70% of the observed discharge data (red circles). The strong similarity between the simulation results of the calibration and evaluation sample inspires confidence in the ability of the behavioral parameter set to accurately describe the rainfall-runoff transformation of the French Broad river basin.

Table 3.4 summarizes the results of GLUE and ABC for the SAC-SMA model and French Broad watershed, and presents the root mean squared error (RMSE) of the posterior mean discharge simulation and associated coverage of the 95% prediction intervals for the calibration and evaluation period. For completeness, we also list the results of the SAC-SMA model with a formal likelihood function, Equation (3.4) using the heteroscedastic measurement error, $\hat{\boldsymbol{\sigma}}_{\tilde{\mathbf{Y}}}$ derived from the nonparametric difference operator. The listed statistics summarize our main findings thus far. The results of GLUE and ABC are virtually identical and show a consistent performance during the calibration and evaluation period. The 95% uncertainty ranges derived with both methods encompass about 70% of the discharge observations. This

coverage of the parameter uncertainty is significantly larger than the approximately 12-17% derived from a classical likelihood function. This finding has important practical utility, for instance within the context of flood forecasting. The behavioral parameter distribution derived with ABC and GLUE provides a reasonable initial estimate of the total out-of-sample prediction uncertainty. On the contrary, the posterior parameter uncertainty derived from a classical likelihood function only envelops a small percentage of the streamflow observations. It is worth noting that this coverage will be inflated if other sources of uncertainty are considered in the formal Bayesian analysis. Yet, this requires a better understanding of precipitation and model structural errors.

Table 3.4: Summary statistics of the GLUE, ABC and DREAM (formal likelihood function) derived posterior parameter distribution for the calibration and evaluation period of the French Broad river basin: Root mean square error (RMSE) of the posterior mean SAC-SMA simulation and coverage of the associated 95% streamflow simulation uncertainty ranges.

	ABC		DREAM		GLUE	
	Calibration	Evaluation	Calibration	Evaluation	Calibration	Evaluation
RMSE, m^3/s	5.44	4.76	4.79	4.85	5.33	4.63
Coverage, %	73.89	69.89	17.35	11.59	76.74	71.08

Our main focus thus far has been on the SAC-SMA model, without recourse to the simulation results of the hmodel. Figure 3.9 shows posterior histograms of five of the hmodel parameters derived with GLUE (top panel) and ABC (bottom panel) using the French Broad calibration data set. The PMC sampler determined a maximum possible coverage of 95% of the discharge data within the uniform hypercube defined by the effective observation error. The results in Figures 3.9-3.11 thus pertain to this coverage level. The marginal posterior parameter distributions derived with GLUE and ABC again demonstrate a strong agreement. Most of the hmodel parameters, with the exception of $I_{\max}$ and α_E are reasonably well defined by calibration against the observed discharge data. The parameter $Q_{\max}$ is particularly well resolved and favors values close to zero - something that is physically rather unrealistic and likely due to errors in the model formulation and precipitation data.

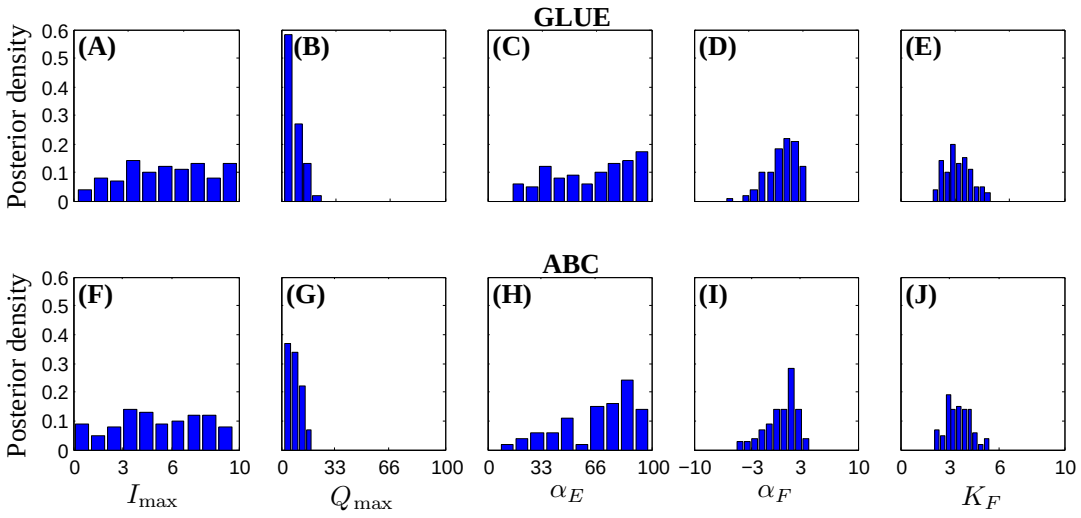


Figure 3.9: Posterior distribution of five randomly selected hmodel parameters derived from the French Broad calibration data set, (A) $I_{\max}$, (B) $Q_{\max}$, (C) α_E, (D) α_F, and (E) K_F.

Figure 3.10 presents two-dimensional scatter plots of the posterior samples of three selected parameter pairs. The top panel corresponds to GLUE and the bottom panel illustrates the same results for ABC. Each plus symbol depicts a behavioral solution. Because of sampling inefficiency the GLUE calculations were terminated after 100 behavioral samples were identified. This explains the apparent differences in sampling density. Nevertheless, the bivariate plots of the posterior samples derived with both methods are in strong agreement with each other, with behavioral solutions that occupy only a relatively small part of the prior parameter space. This is particularly true for the $\alpha_F - Q\text{max}$ subspace. The sampled parameter pairs appear rather uncorrelated which suggests that the different hmodel parameters each control a different part of the simulated watershed response. This simplifies posterior inference, and favors a hierarchical sampling approach in which parameters are estimated sequentially.

We now demonstrate how the hmodel parameter uncertainty translates into streamflow simulation uncertainty. We separately depict the results for the calibration (Figure 3.11) and evaluation (Figure 3.12) period. As expected, the simulation results derived with GLUE and ABC are in close agreement. The 95% simulation uncertainty ranges encompass about

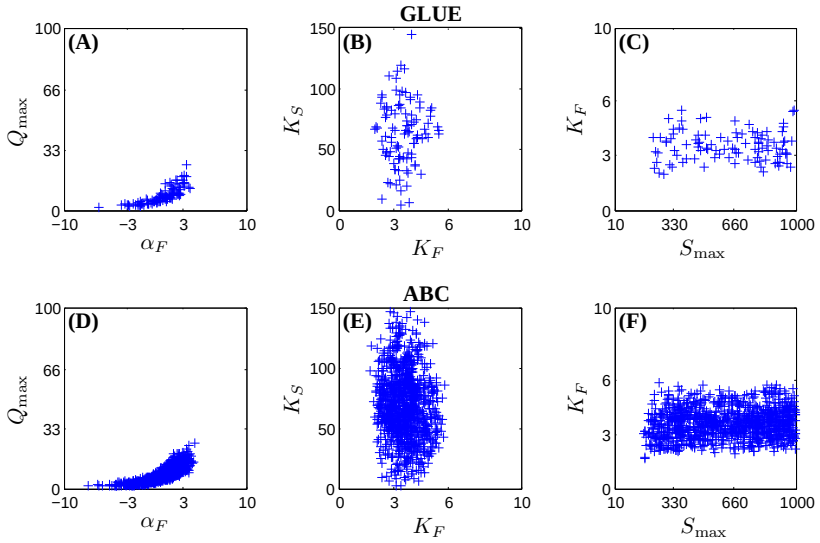


Figure 3.10: Bivariate scatter plots of the behavioral (posterior) samples of three different (randomly selected) hmodel parameter pairs using GLUE (top panel) and ABC (bottom panel): (A) $\alpha_F - Q_{\max}$, (B) $K_F - K_S$, and (C) $S_{\max} - K_F$. The scatter plots derived with both methods are in close agreement but demonstrate an important difference in sampling density. The computational budget of GLUE was limited to about 3 days, and this has resulted in 100 behavioral solutions.

Table 3.5: Summary statistics of the GLUE, ABC and DREAM (formal likelihood function) derived posterior parameter distribution for the calibration and evaluation period of the French Broad river basin: Root mean square error (RMSE) of the posterior mean hmodel simulation and coverage (%) of the associated 95% streamflow simulation uncertainty ranges.

	ABC		DREAM		GLUE	
	Calibration	Evaluation	Calibration	Evaluation	Calibration	Evaluation
RMSE, m^3/s	6.14	5.80	5.23	5.08	6.15	5.76
Coverage, %	67.38	61.86	7.33	10.58	68.14	60.86

67% of the calibration data observations (see Table 3.5). This is much higher than the 7% coverage derived with a classical likelihood function using MCMC simulation with DREAM. Yet, between days 205 and 270 the posterior ensemble systematically over predicts the actual discharge observations. This positive bias is likely caused by an error in the measured rainfall data around day 205. This rainfall error accumulates in the simulated state variables and continues to persist until the next significant rainfall event around day 270. Rainfall data correction would seem appropriate [*Kavetski et al.*, 2006a,b, *Vrugt et al.*, 2008a, *Beven*, 2009], but is beyond the scope of the present chapter.

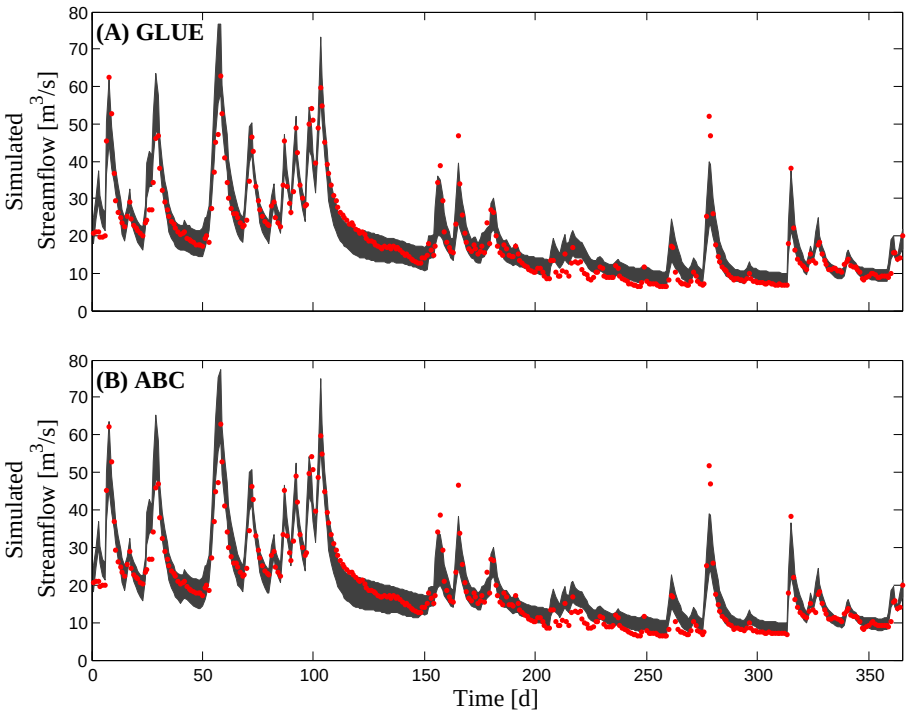


Figure 3.11: Hmodel derived 95% streamflow simulation uncertainty ranges (grey region) of the calibration period of the French Broad river basin using (A) GLUE and (B) ABC. The observed discharge data are indicated with the solid red dots. We limit our display to the first 365-days of the calibration data set to facilitate graphical interpretation. The uncertainty ranges appear very similar and nicely cover the observed discharge data.

The evaluation data period again highlights the strong operational similarity of GLUE and ABC, but the average width of the 95% streamflow simulation uncertainty ranges appears somewhat smaller. Indeed, the coverage has reduced to approximately 61%. Note that the posterior ensemble systematically underestimates the peak flow events. This can be the effect of an increased rainfall intensity during the evaluation period or some sort of epistemic error [*Beven*, 2006, 2009, *Beven et al.*, 2011]. For practical application it would seem most productive to extend the length of the calibration data period to include a number of larger storm events. This would certainly improve the fitting of the peak flow events, but not affect the main conclusions of this chapter.

We now turn our attention on the Leaf River data set, and present in Figure 3.13 histograms of the marginal posterior distributions of a representative group of five SAC-SMA param-

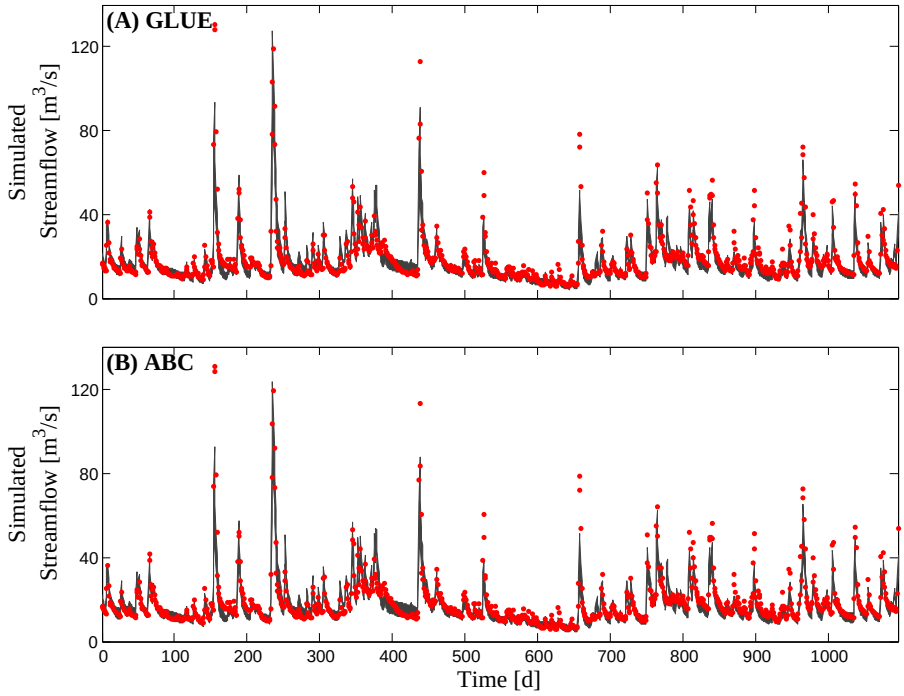


Figure 3.12: Hmodel derived 95% streamflow simulation uncertainty ranges (grey region) for a three-year portion of the evaluation period of the French Broad river basin using the (A) GLUE and (B) ABC derived posterior parameter distribution. The observed discharge data are indicated with the solid red dots.

eters. The PMC sampler determined a maximum coverage of about 56% of the discharge data within the effective observation error used herein. The top panel displays the results for GLUE for this coverage level, whereas the bottom panel shows the corresponding counterparts derived with ABC. The histograms derived with both method again are strikingly similar, yet the PMC sampler used in ABC is about 30 times more efficient (not tabulated) in sampling the $N = 1,000$ behavioral solutions. Note that SAC-SMA parameters are not particularly well defined by calibration against the limits of acceptability. The marginal posterior uncertainty ranges are rather large, with the exception of the parameter LZPK that tends to favor values close to zero.

Finally, Figure 3.14 illustrates the performance of the GLUE and ABC derived posterior parameter distributions for an independent evaluation period. The GLUE (top panel) and ABC (bottom panel) derived 95% posterior streamflow uncertainty ranges again appear

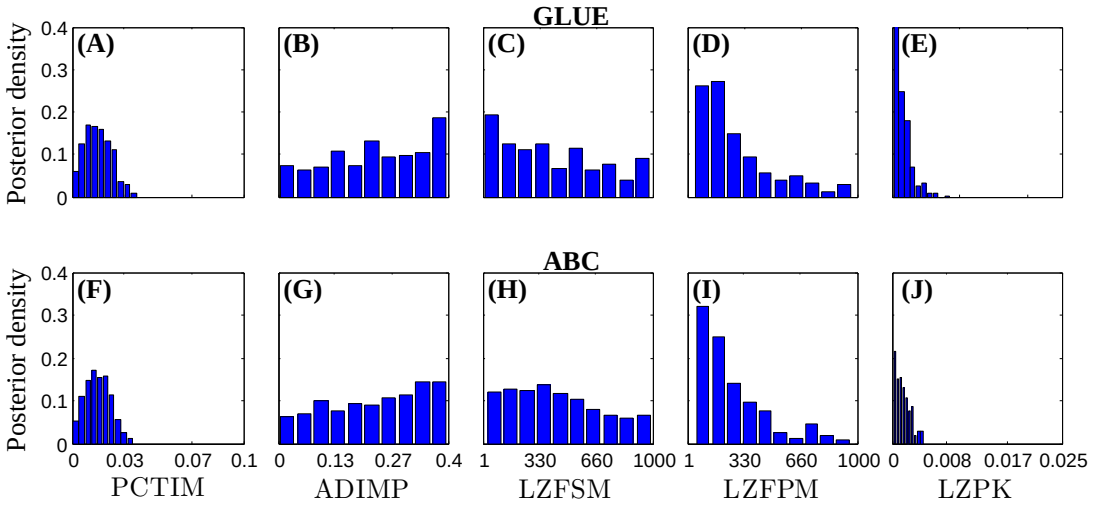


Figure 3.13: Posterior distribution of five randomly chosen SAC-SMA model parameters derived from GLUE (top panel), and ABC (bottom panel) using historical streamflow data from the Leaf River watershed in Mississippi.

nearly equivalent (expected) and contain about 65% of the observed discharge values. This is somewhat smaller than the 80% coverage obtained during the calibration data period (not shown herein).

3.6 Discussion and Conclusions

In the past two decades, the GLUE methodology of *Beven and Binley* [1992], *Beven and Freer* [2001], *Beven* [2006] has found widespread application and use for model parameter and predictive uncertainty analysis. This method rejects the formal Bayesian paradigm in favor of finding a set of of behavioral solutions that are acceptably close to the non-error-free observations. This avoids over conditioning of the posterior parameter space in non-ideal cases with nontraditional error residual distributions. Indeed, Tables 3.4 - 3.6 demonstrate that the GLUE derived 95% simulation uncertainty ranges encompass a much higher percentage of the discharge observations than the posterior parameter predictive uncertainty intervals derived from a classical likelihood function. Formal likelihood functions that do

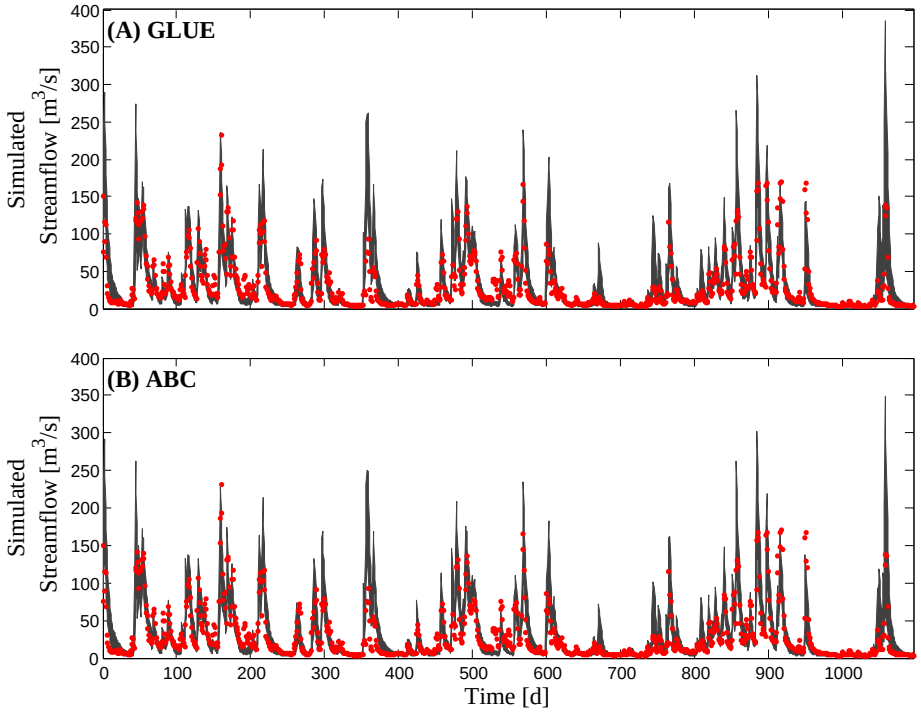


Figure 3.14: SAC-SMA derived 95% streamflow simulation uncertainty ranges (grey region) for a three-year portion of the evaluation period of the Leaf River watershed using the (A) GLUE and (B) ABC derived posterior parameter distribution. The observed discharge data are indicated with the solid red dots.

not adequately describe the probabilistic properties of the error residuals tend to overestimate the actual information content of the data, and provide estimates of the posterior parameter uncertainty that are overly optimistic.

Many have criticized the GLUE methodology for being subjective and lacking an appropriate mathematical underpinning. To help bridge the gap between informal and formal Bayesian approaches, this chapter introduced likelihood-free inference to hydrologic modeling and uncertainty analysis. This approach was introduced in the statistical literature about three decades ago [*Diggle and Gratton*, 1984] for cases when an explicit likelihood (objective) function cannot be justified. Such approaches, also referred to as ABC, use one or multiple (sufficient) statistics to estimate the posterior parameter distribution. The premise behind ABC is that $\boldsymbol{\theta}'$ should be a sample from the posterior distribution as long as the distance between the observed and simulated summary statistics is smaller than some small value, ϵ.

Table 3.6: Summary statistics of the GLUE, ABC and DREAM (formal likelihood function) derived posterior parameter distribution for the calibration and evaluation period of the Leaf River watershed: Root mean square error (RMSE) of the posterior mean SAC-SMA simulation and coverage (%) of the associated 95% streamflow simulation uncertainty ranges.

	ABC		DREAM		GLUE	
	Calibration	Evaluation	Calibration	Evaluation	Calibration	Evaluation
RMSE, m^3/s	20.95	20.76	16.39	19.23	20.73	20.62
Coverage, %	80.02	64.78	22.11	21.90	82.54	68.25

An example of this was given in Section 3.3 by calibration of the Nash-Cascade model against the mean and standard deviation of the discharge data. In the limit of ϵ going to zero, the behavioral solution space should converge to the actual posterior distribution, pending the assumption that the chosen summary statistic(s) is (are) near sufficient [*Pritchard et al.*, 1999, *Vrugt and Sadegh*, 2013]. But this was certainly not the case for the Nash-Cascade example. The mean and standard deviation are rather weak summary metrics and this explains why the marginal posterior distribution of the recession constant was too wide (see Figure 3.3C) and did not converge to its expected distribution (Figure 3.3B). Thus, there is a clear need for meaningful summary statistics with a compelling diagnostic power. Examples include the annual runoff and baseflow coefficient, and the flow duration curve as used in *Vrugt and Sadegh* [2013]. Note that the ABC approach differs from multiple objective calibration methods in that the distance between the observed and simulated summary metrics is jointly minimized.

Numerical simulations presented in Figs. 3.4-3.14 have shown that, if each observation is treated as a summary variable, then the ABC approach obtains very similar results as the limits of acceptability approach of GLUE. A similar conclusion was drawn in previous work by *Nott et al.* [2012] but following a different line of reasoning and within the context of the more traditional GLUE methodology presented by *Beven and Binley* [1992]. One issue deserves special attention, and that is that within the limit of acceptability framework, the value of ϵ needs to be taken much larger than what is deemed statistically adequate.

Standard applications of likelihood free inference define a solution to be behavioral if the chosen summary statistics are within a small distance of their observed counterparts. Yet for hydrologic systems with many calibration observations the probability of happening upon a simulation run that yields exactly the same data set as the one observed will be extremely small. The effective observation error remedies this problem, but the magnitude of this value is typically much larger than the theoretical value of ϵ to guarantee converge to the true posterior parameter distribution. Thus, although the limits of acceptability approach of GLUE is a special variant of the more generic ABC approach, one should be careful with interpretation of the posterior parameter distribution.

The results in Tables 3.4 - 3.6 demonstrate that the RMSE of the posterior mean ABC (GLUE) simulation is substantially larger than its counterpart derived with DREAM using a formal likelihood function. This finding is not surprising. The likelihood function used in DREAM is specifically designed to minimize the squared distance between the model simulation and corresponding data. This metric poses much stronger constraints on the parameter values than the limits of acceptability used in GLUE or ABC, and hence results in a better compliance of the simulated and observed discharge data. The performance of the posterior mean ABC simulation can be enhanced if the limits of acceptability, $\boldsymbol{\delta}$ are tightened and/or additional summary metrics are used during model fitting. Ideally, the chosen summary statistic, $S(\cdot)$ is sufficient for $\boldsymbol{\theta}$ and thus provides as much information for the parameters as the original data set, $\tilde{\mathbf{Y}}$ itself. However, if the exact (perfect) likelihood function is unknown, it will be difficult to determine a sufficient statistic. One could then use multiple different summary statistics that each capture different aspects/signatures/patterns of the input-output response [*Ratmann et al.*, 2009, *Vrugt and Sadegh*, 2013].

The findings presented thus far are derived by comparing the SAC-SMA and hmodel simulated streamflow time series with the observed data. This deterministic approach does not generate a random sample at each time step, and therefore violates a basic requirement of

ABC. Nevertheless, the use of a deterministic model is adequate within the current context, because the limits of acceptability used herein are substantially larger (four times) than the measurement error of the streamflow data. To illustrate this, please consider Figure 3.15 that plots the ABC-derived parameter and 95% posterior simulation uncertainty ranges for the French Broad river basin using deterministic (top panel) and stochastic (bottom panel) SAC-SMA modeling. We limit our display to a 250-day portion of the calibration data period, and the marginal posterior distribution of two selected SAC-SMA model parameters (right-hand side). The stochastic simulation is simply derived by adding a random measurement error to the daily modeled SAC-SMA discharge values. As expected, the posterior streamflow uncertainty ranges and posterior histograms appear very similar. This concludes our numerical simulations.

The ABC methodology opens an entire new field of research with infinite scope for (amongst others) (a) inventing new summary metrics that are self-sufficient and properly rooted in hydrologic and information theory, (b) developing methods and guidelines to interpret the outcome of the ABC analysis and help aid in detection of model malfunctioning [*Vrugt and Sadegh*, 2013], (c) exploring ways to determine the temporal/spatial variability of individual summary metrics for stochastic (Bayesian) analysis, (d) new approaches to diagnostic model selection using stochastic minimization of the mutual information, (e) improving the computational efficiency of ABC-sampling [*Sadegh and Vrugt*, 2013], and (f) practical applications. Challenges lie in the proper selection of summary metrics that properly extract the information from the calibration data, how to deal with input data uncertainty, and how to detect epistemic errors (lack of knowledge).

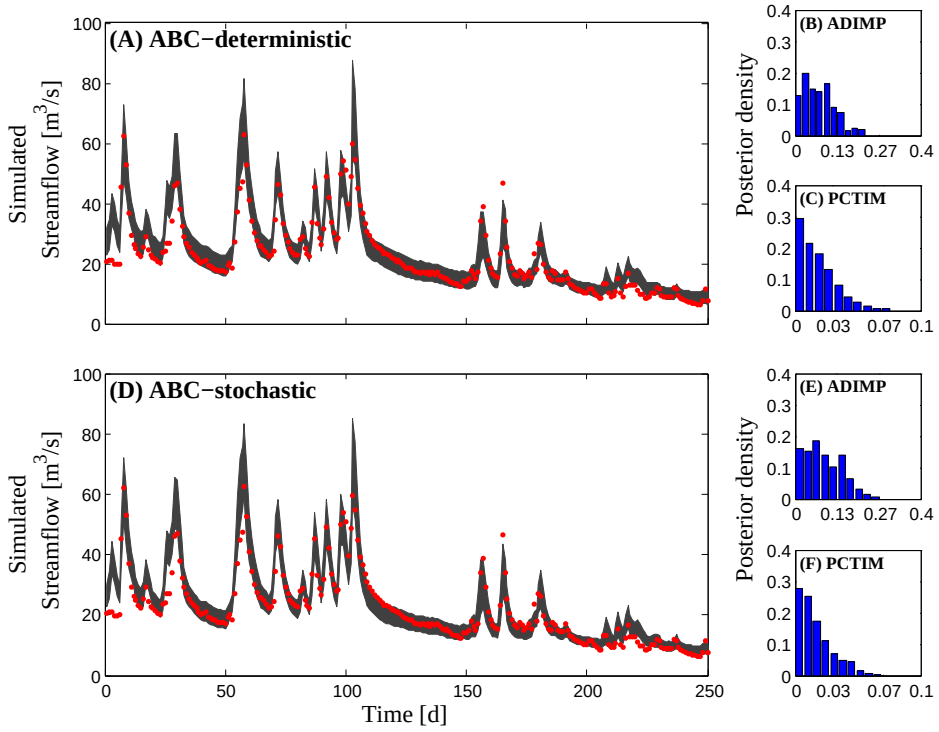


Figure 3.15: Illustration of the effect of measurement data uncertainty on the ABC-derived posterior parameter and streamflow uncertainty for a representative 250-day portion of the French Broad calibration data set: (A) 95% simulation uncertainty assuming a deterministic, $\mathbf{Y} \leftarrow \mathcal{H}(\boldsymbol{\theta}|\cdot)$, and (D) stochastic model operator, $\mathbf{Y} \leftarrow \mathcal{H}(\boldsymbol{\theta}|\cdot) + N(0, \hat{\sigma}_{\tilde{\mathbf{Y}}})$ with discharge measurement error derived from the nonparametric estimator in equation (3.11). The observed discharge data are indicated with the solid red dots. The histograms at the right-hand side of each individual panel plot the corresponding marginal posterior distributions of the SAC-SMA model parameters ADIMP and PCTIM.

3.A Appendix A

Suppose some measurement data $\tilde{\mathbf{Y}} = \{\tilde{y}_1, \ldots, \tilde{y}_n\}$, and a model that predicts $\mathbf{Y} \leftarrow \mathcal{H}(\theta|\cdot)$ with parameter values, $\boldsymbol{\theta} \in \boldsymbol{\Theta} \in \mathbb{R}^d$. We define a prior distribution, $p(\boldsymbol{\theta})$ 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]

This concludes the pseudo-code of the population Monte Carlo sampler.

Algorithm 5 ABC-PMC

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

3.B Appendix B

The PMC sampler of *Turner and van Zandt* [2012] assumes that the sequence of ϵ values is specified by the user. Practical experience suggests that a poor selection of $\varepsilon = \{\epsilon_1, \ldots, \epsilon_J\}$ can lead to very low acceptance rates or even premature convergence if ϵ has been taken too small. We therefore implement an arguably more advanced strategy and let the sampler adaptively select the values of $\epsilon_j (j > 1)$.

This strategy is implemented as follows. The user defines ϵ_1. In practice, a large value will typically suffice. At the end of the first iteration (just after σ_1^2 has been calculated), the algorithm computes the cumulative distribution function (cdf) of the N accepted $\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y})\big)$ values. This function ranges between 0 and 1 and describes the probability that a random variable X (in this case $\rho\big(\cdot\big)$) will be found at a value less than or equal to x. The value of ϵ_2 is then taken to be that value of $\rho\big(\cdot\big)$ at which the cdf is equal to 0.1. The PMC sampler proceeds with the next iteration, $j = 2$ and this recipe is continued during each successive next iteration until ϵ_j reaches some lower default value defined by the user (Section 3.3), or when the successive reduction in ϵ has become smaller than 0.02; $(\epsilon_j - \epsilon_{j-1} < 0.02)$ (Section 3.4).

Numerical simulation has shown that this adaptive updating strategy of ϵ significantly enhances the search efficiency of the PMC sampler. Moreover, this implementation does not require the user to specify $\varepsilon = \{\epsilon_2, \ldots, \epsilon_J\}$, which is an important practical advantage.

Chapter 4

Approximate Bayesian Computation Using Markov Chain Monte Carlo Simulation: DREAM$_{(\mathrm{ABC})}$

This Chapter has been published in Water Resources Research. Citation: Sadegh, M., and J.A. Vrugt (2014), Approximate Bayesian Computation using Markov Chain Monte Carlo simulation: DREAM$_{(\mathrm{ABC})}$, *Water Resources Research*, *50*, 67676787, doi:10.1002/2014WR015386.

4.1 Introduction and Scope

Bayesian methods have become increasingly popular for fitting hydrologic models to data (e.g. streamflow, water chemistry, groundwater table depth, soil moisture, pressure head, snow water equivalent). Bayes' rule updates the prior probability of a certain hypothesis when new data, $\tilde{\mathbf{Y}} = \{\tilde{y}_1, \ldots, \tilde{y}_n\}$ (also referred to as evidence) become available. The hypothesis typically constitutes the parameter values, $\boldsymbol{\theta}$, of a model, F, which simulates the observed data using

$$\tilde{\mathbf{Y}} \leftarrow F(\boldsymbol{\theta}, \tilde{\mathbf{u}}, \tilde{\mathbf{x}}_0) + \mathbf{e}, \tag{4.1}$$

where $\tilde{\mathbf{u}} = \{\tilde{u}_1, \ldots, \tilde{u}_n\}$ denotes the observed forcing data, $\tilde{\mathbf{x}}_0$ signifies the initial state of the system, and $\mathbf{e} = \{e_1, \ldots, e_n\}$ includes observation error, as well as error due to the fact that the simulator, F, may be systematically different from the real system of interest, $\Im(\boldsymbol{\theta})$, for the parameters $\boldsymbol{\theta}$. If our main interest is in the parameters of the model, Bayes law is given by

$$p(\boldsymbol{\theta}|\tilde{\mathbf{Y}}) = \frac{p(\boldsymbol{\theta})p(\tilde{\mathbf{Y}}|\boldsymbol{\theta})}{p(\tilde{\mathbf{Y}})}, \tag{4.2}$$

where $p(\boldsymbol{\theta})$ denotes the prior parameter distribution, $p(\tilde{\mathbf{Y}}|\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$ is the likelihood function, and $p(\tilde{\mathbf{Y}})$ represents the evidence. As all statistical inferences of the parameters can be made from the unnormalized density, we conveniently remove $p(\tilde{\mathbf{Y}})$ from the denominator and write $p(\boldsymbol{\theta}|\tilde{\mathbf{Y}}) \propto p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$.

The likelihood function, $L(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$, summarizes, in probabilistic sense, the overall distance between the model simulation and corresponding observations. The mathematical definition of this function has been subject to considerable debate in the hydrologic and statistical literature [e.g. *Schoups and Vrugt* [2010a], *Smith et al.* [2010], *Evin et al.* [2013]]. Simple

likelihood functions that assume Gaussian error residuals are statistically convenient, but this assumption is often not borne out of the probabilistic properties of the error residuals that show significant variations in bias, variance, and autocorrelation at different parts of the simulated watershed response. Such non-traditional residual distributions are often caused by forcing data and model structural errors, whose probabilistic properties are very difficult, if not impossible, to adequately characterize. This makes it rather difficult, if not impossible, to isolate and detect epistemic errors (model structural deficiencies), a prerequisite to improving our understanding and theory of water flow and storage in watersheds.

The inability of classical likelihood-based fitting methods to detect model malfunctioning is evident if we critically assess the progress that has been made in modeling of the rainfall-runoff transformation. For instance, consider the Sacramento soil moisture accounting (SAC-SMA) model introduced by *Burnash et al.* [1973] in the early 1970s and used by the US National Weather Service for flash-flood forecasting throughout the United States. In about four decades of fitting the SAC-SMA model to (spatially distributed) streamflow data, we have not been able to make any noticeable improvements to the underlying equations of the model. This is even more disturbing given the relative low order complexity of the SAC-SMA model. If for such relatively simple (lumped) hydrologic models our fitting methods are unable to illuminate to what degree a representation of the real world has been adequately achieved and how the model should be improved, the prospects of learning and scientific discovery for the emerging generation of very high order system models are rather poor, because more complex process representations lead (unavoidably) to greater interaction among model components, and perpetually larger volumes of field and remote sensing data need to be utilized for system characterization and evaluation.

The limitations of classical residual-based fitting methods has stimulated *Gupta et al.* [2008] (amongst others) to propose a signature-based approach to model evaluation. By choosing the signatures so that they each measure different but relevant parts of system behavior,

diagnostic evaluation proceeds with analysis of the behavioural similarities (and differences) of the observed data and corresponding model simulations. Ideally, these differences are then related to individual process descriptions, and correction takes place by refining/improving these respective components of the model. What is left is the numerical implementation of diagnostic model evaluation.

In Chapter 2 we advocated the use of "likelihood-free" inference for diagnostic model evaluation. This approach, introduced in the statistical literature about three decades ago [*Diggle and Gratton*, 1984], is especially useful for cases where the likelihood is intractable, too expensive to be evaluated, or impossible to be formulated explicitly. This class of methods is also referred to as Approximate Bayesian Computation (ABC), a term coined by *Beaumont et al.* [2002], and widens the realm of models for which statistical inference can be considered [*Marjoram et al.*, 2003, *Sisson et al.*, 2007, *Del Moral et al.*, 2008, *Joyce and Marjoram*, 2008, *Grelaud et et al.*, 2009, *Ratmann et al.*, 2009]. ABC has rapidly gained popularity in the past few years, in particular for the analysis of complex problems arising in population genetics, ecology, epidemiology, and systems biology. The first application of ABC in hydrology can be found in *Nott et al.* [2012] and establishes a theoretical connection between ABC and GLUE-based approaches. Other work on this topic can be found in the recent publication by *Sadegh and Vrugt* [2013].

The premise behind ABC is that $\boldsymbol{\theta}^*$ should be a sample from the posterior distribution if the distance between the observed and simulated data, $\rho\big(\tilde{\mathbf{Y}}, \mathbf{Y}(\boldsymbol{\theta}^*)\big)$, is smaller than some small positive value, ϵ [*Marjoram et al.*, 2003, *Sisson et al.*, 2007]. Figure 4.1 provides a conceptual overview of the ABC methodology. All ABC based methods approximate the likelihood function by simulations, the outcomes of which are compared with the observed data [*Beaumont*, 2010, *Bertorelle et al.*, 2010, *Csilléry et al.*, 2010]. In so doing, ABC

algorithms attempt to approximate the posterior distribution by sampling from

$$p(\boldsymbol{\theta}|\tilde{\mathbf{Y}}) \propto \int_{\boldsymbol{\mathcal{Y}}} p(\boldsymbol{\theta})\text{Model}(\boldsymbol{y}|\boldsymbol{\theta})\text{I}(\rho\big(\tilde{\mathbf{Y}}, \mathbf{Y}(\boldsymbol{\theta})\big) \leq \epsilon)d\boldsymbol{y}, \tag{4.3}$$

where $\boldsymbol{\mathcal{Y}}$ denotes the support of the simulated data, $\mathbf{Y} \sim \text{Model}(\boldsymbol{y}|\boldsymbol{\theta})$, and $\text{I}(a)$ is an indicator function that returns one if the condition a is satisfied and zero otherwise. The accuracy of the estimated posterior distribution, $p(\theta|\rho\big(\tilde{\mathbf{Y}}, \mathbf{Y}(\boldsymbol{\theta})\big) \leq \epsilon)$ depends on the value of ϵ. In the limit of $\epsilon \to 0$ the sampled distribution will converge to the true posterior, $p(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$ [*Pritchard et al.*, 1999, *Beaumont et al.*, 2002, *Ratmann et al.*, 2009, *Turner and van Zandt*, 2012]. Yet, this requires the underlying model operator to be stochastic, and hence Model$(\cdot)$ in Equation (4.3) is equivalent to the output of the deterministic model F in Equation (4.1) plus a random error with probabilistic properties equal to those of $\mathbf{e}$.

For sufficiently complex system models and/or large data sets, it will be difficult, if not impossible, to find a model simulation that always fits the data within ϵ. It has therefore become common practice in ABC to use one or more summary statistics of the data rather than the data itself. Ideally, these chosen summary statistics, $S(\cdot)$ are sufficient and thus provide as much information for the model parameters as the original data set itself. In practice, however, the use of summary statistics usually entails a loss of information and hence results in an approximate likelihood, especially for complex models. Partial least squares [*Wegmann et al.*, 2009] and information-theory [*Barnes et al.*, 2011] can help to determine (approximately) a set of nearly-sufficient marginal statistics. Nonetheless, complex models admitting sufficient statistics are practical exceptions.

The most common ABC algorithm implements simple rejection sampling which relies on satisfying the condition $\rho\big(S((\tilde{\mathbf{Y}})), S(\mathbf{Y}(\boldsymbol{\theta}^*))\big) \leq \epsilon$. This method has the practical advantage of being relatively easy to implement and use, but its efficiency depends critically on the choice of the prior sampling distribution. If this prior distribution is a poor approximation of

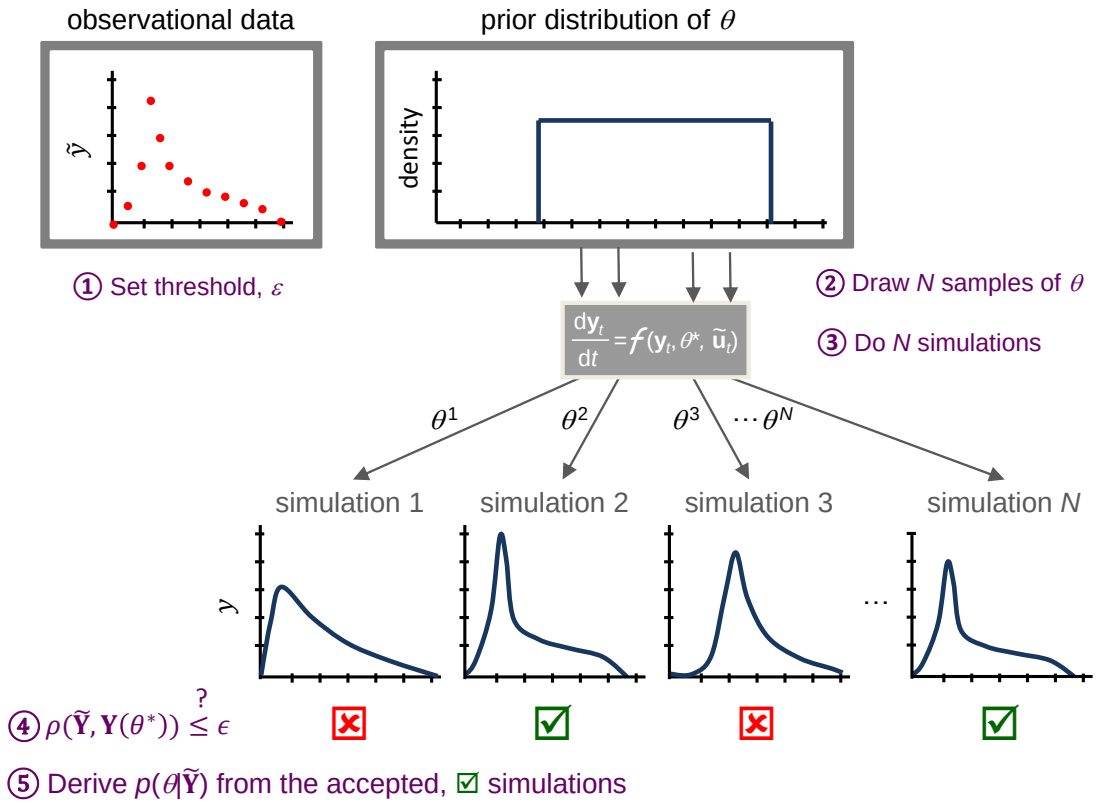


Figure 4.1: Conceptual overview of approximate Bayesian computation (ABC) for a hypothetical one-dimensional parameter estimation problem. First, N samples are drawn from a user-defined prior distribution, $\theta^* \sim p(\theta)$. Then, this ensemble is evaluated with the model and creates N model simulations. If the distance between the observed and simulated data, $\rho(\tilde{\mathbf{Y}}, \mathbf{Y}(\theta^*))$ is smaller than or equal to some nominal value, ϵ then θ^* is retained, otherwise the simulation is discarded. The accepted samples are then used to approximate the posterior parameter distribution, $p(\theta|\tilde{\mathbf{Y}})$. Note that for sufficiently complex models and large data sets the probability of happening upon a simulation run that yields precisely the same simulated values as the observations will be very small, often unacceptably so. Therefore, $\rho(\tilde{\mathbf{Y}}, \mathbf{Y}(\boldsymbol{\theta}^*))$ is typically defined as a distance between summary statistics of the simulated, $S(\mathbf{Y}(\boldsymbol{\theta}^*))$ and observed, $S(\tilde{\mathbf{Y}})$ data, respectively. Modified after *Sunnåker et al.* [2013].

the actual posterior distribution, then many of the proposed samples will be rejected. This leads to dramatically low acceptance rates, and thus excessive CPU-times. Indeed, *Vrugt and Sadegh* [2013] and *Sadegh and Vrugt* [2013] report acceptance rates of less than 0.1% for hydrologic models with just a handful of parameters. One remedy to this problem is to increase the value of ϵ, but this leads to an inaccurate approximation of the posterior distribution.

A number of methodological advances have been proposed to enhance the sampling efficiency

of ABC algorithms. One common approach is to use a set of monotonically decreasing ϵ values. This allows the algorithm to sequentially adapt the prior distribution and converge to a computationally feasible final value of ϵ. Nonetheless, these algorithms still rely on a boxcar kernel (step function) to evaluate the fitness of each sample, and are not particularly efficient in high dimensional search spaces. In this chapter we introduce a Markov Chain Monte Carlo (MCMC) simulation method that enhances, sometimes dramatically, the ABC sampling efficiency. This general-purpose method entitled, DREAM$_{(\mathrm{ABC})}$ uses the DiffeRential Evolution Adaptive Metropolis algorithm [*Vrugt et al.*, 2008a, 2009a] as its main building block, and replaces the indicator function in Equation (4.3) with a continuous kernel to decide whether to accept candidate points or not. The proposed methodology is benchmarked using synthetic and real-world simulation experiments.

The remainder of this chapter is organized as follows. In section 4.2 we summarize the results of commonly used ABC sampling methods by application to a synthetic benchmark study. Section 4.3 introduces the main elements of the DREAM$_{(\mathrm{ABC})}$ algorithm and discusses several of its advantages. This is followed in Section 4.4 with two synthetic and one real-world simulation experiment. In this section we are especially concerned with sampling efficiency and robustness. Finally, section 4.5 concludes this chapter with a discussion and summary of our main findings.

4.2 Approximate Bayesian Computation

The ABC method provides an excellent vehicle for diagnostic model evaluation [*Vrugt and Sadegh*, 2013] by using one or multiple different summary statistics that, when rooted in the relevant environmental theory, should have a much stronger and compelling diagnostic power than some residual-based likelihood function. Challenges lie in the proper selection of summary metrics that adequately extract all the available information from the calibra-

tion data, how to deal with input data uncertainty, how to detect epistemic errors (lack of knowledge), how to determine an appropriate (small) value for ϵ, and how to efficiently sample complex multi-dimensional spaces involving many tens to hundreds of parameters. This chapter is focused on the last topic, and proposes several methodological developments to overcome the shortcomings of standard ABC sampling methods.

We first discuss two common ABC sampling methods that have found widespread application and use within the context of likelihood-free inference. We then introduce DREAM$_{(\text{ABC})}$, a Markov chain Monte Carlo (MCMC) implementation of ABC that permits inference of complex system models.

4.2.1 Rejection Algorithm

Once the summary statistic(s) has(have) been defined we are left with finding all those values of $\boldsymbol{\theta}^*$ for which $\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}^*))\big) \leq \epsilon$. The most basic algorithm to do so uses rejection sampling. This algorithm proceeds as follows

Algorithm 6 ABC-Rejection Sampler

for $i = 1, \ldots, N$ **do**
  while $\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}^*))\big) > \epsilon$ **do**
    Sample $\boldsymbol{\theta}^*$ from the prior, $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta})$
    Simulate data $\mathbf{Y}$ using $\boldsymbol{\theta}^*$, $\mathbf{Y} \sim \text{Model}(\boldsymbol{\theta}^*)$
    Calculate $\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}^*))\big)$
  end while
  Set $\boldsymbol{\theta}^i \leftarrow \boldsymbol{\theta}^*$
  Set $w^i \leftarrow \frac{1}{N}$
end for

In words, the ABC rejection (ABC-REJ) algorithm proceeds as follows. First we sample a candidate point, $\boldsymbol{\theta}^*$, from some prior distribution, $p(\boldsymbol{\theta})$. We then use this proposal to simulate the output of the model, $\mathbf{Y} \sim \text{Model}(\boldsymbol{\theta}^*)$. We then compare the simulated data, $\mathbf{Y}$, with the observed data, $\tilde{\mathbf{Y}}$, using a distance function, $\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}^*))\big)$. If this distance

function is smaller than some small positive tolerance value, ϵ then the simulation is close enough to the observations that the candidate point, $\boldsymbol{\theta}^*$ has some nonzero probability of being in the approximate posterior distribution, $\hat{p}(\boldsymbol{\theta}|\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}))\big) \leq \epsilon)$. By repeating this process N times, ABC-REJ provides an estimate of the actual posterior distribution.

Unfortunately, standard rejection sampling method typically requires massive computational resources to generate a sufficient number of samples from the posterior distribution. Failure to maintain an adequate sampling density may result in under sampling probable regions of the parameter space. This inefficiency can provide misleading results, particularly if $p(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$ is high dimensional and occupies only a small region interior to the prior distribution. Only if $p(\boldsymbol{\theta})$ is a good representation of the actual posterior parameter distribution then ABC-REJ can achieve an adequate sampling efficiency.

4.2.2 Population Monte Carlo Simulation

To guarantee convergence to the appropriate limiting distribution, the value of ϵ in Algorithm 6 (ABC-REJ) needs to be taken very small. Values of $0.01 \leq \epsilon \leq 0.05$ are often deemed appropriate. Unfortunately, this will produce very low acceptance rates, particularly if the prior distribution is poorly chosen and extends far beyond the posterior distribution. To increase sampling efficiency, it would seem logical to stepwise reduce the value of ϵ and to use the accepted samples to iteratively adapt the prior distribution. This is the principal idea behind population Monte Carlo (PMC) algorithms. These methods are used extensively in physics and statistics for many-body problems, lattice spin systems and Bayesian inference, and also referred to as "quantum Monte Carlo", "transfer-matrix Monte Carlo", "Monte Carlo filter", "particle filter" and "sequential Monte Carlo". The PMC sampler of *Beaumont et al.* [2009], *Turner and van Zandt* [2012] is specifically designed for ABC-inference and works as follows

Algorithm 7 ABC-Population Monte Carlo sampler

At iteration $j = 1$,
for $i = 1, \ldots, N$ **do**
  while $\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}^*))\big) > \epsilon_1$ **do**
    Sample $\boldsymbol{\theta}^*$ from the prior, $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta})$
    Simulate data $\mathbf{Y}$ using $\boldsymbol{\theta}^*$, $\mathbf{Y} \sim \text{Model}(\boldsymbol{\theta}^*)$
    Calculate $\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}^*))\big)$
  end while
  Set $\boldsymbol{\Theta}_1^i \leftarrow \boldsymbol{\theta}^*$
  Set $\mathbf{w}_1^i \leftarrow \frac{1}{N}$
end for
Set $\boldsymbol{\Sigma}_1 \leftarrow 2\text{Cov}(\boldsymbol{\Theta}_1)$,
At iteration $j > 1$,
for $j = 2, \ldots, J$ **do**
  for $i = 1, \ldots, N$ **do**
    while $\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}^{**}))\big) > \epsilon_j$ **do**
      Sample $\boldsymbol{\theta}^*$ from the previous population, $\boldsymbol{\theta}^* \sim \boldsymbol{\Theta}_{j-1}$ with probability $\mathbf{w}_{j-1}$
      Perturb $\boldsymbol{\theta}^*$ by sampling $\boldsymbol{\theta}^{**} \sim \mathcal{N}_d(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}_{j-1})$
      Simulate data $\mathbf{Y}$ using $\boldsymbol{\theta}^{**}$, $\mathbf{Y} \sim \text{Model}(\boldsymbol{\theta}^{**})$
      Calculate $\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}^{**}))\big)$
    end while
    Set $\boldsymbol{\Theta}_j^i \leftarrow \boldsymbol{\theta}^{**}$

    $$\text{Set } w_j^i \leftarrow \frac{p(\boldsymbol{\theta}_j^i)}{\sum_{u=1}^{N} \mathbf{w}_{j-1}^u q_d(\boldsymbol{\theta}_{j-1}^u | \boldsymbol{\theta}_j^i, \boldsymbol{\Sigma}_{j-1})} \tag{4.4}$$

  end for
  Set $\boldsymbol{\Sigma}_j \leftarrow 2\text{Cov}(\boldsymbol{\Theta}_j)$
end for

In short, the ABC-PMC sampler starts out as ABC-REJ during the first iteration, $j = 1$, but using a much larger initial value for ϵ. This will significantly enhance the initial acceptance rate. During each successive iteration, $j = \{2, \ldots, J\}$, the value of ϵ is decreased and the multi-normal proposal distribution, $q = \mathcal{N}_d(\boldsymbol{\theta}_{j-1}^k, \boldsymbol{\Sigma}_{j-1})$, adapted using $\boldsymbol{\Sigma}_{j-1} = 2\text{Cov}(\boldsymbol{\theta}_{j-1}^1, \ldots, \boldsymbol{\theta}_{j-1}^N)$ with $\boldsymbol{\theta}_{j-1}^k$ drawn from a multinomial distribution, $\mathfrak{F}(\boldsymbol{\Theta}_{j-1} | \mathbf{w}_{j-1})$, where $\boldsymbol{\Theta}_{j-1} = \{\boldsymbol{\theta}_{j-1}^1, \ldots, \boldsymbol{\theta}_{j-1}^N\}$ is a $N \times d$ matrix and $\mathbf{w}_{j-1} = \{w_{j-1}^1, \ldots, w_{j-1}^N\}$ is a N-vector of normalized weights, $\sum_{i=1}^{N} w_{j-1}^i = 1$ and $w_{j-1}^i \geq 0$. In summary, a sequence of (multi)normal proposal distributions is used to iteratively refine the samples and explore the posterior

distribution. This approach, similar in spirit as the adaptive Metropolis sampler of [*Haario et al.*, 1999, 2001], achieves a much higher sampling efficiency than ABC-REJ, particularly for cases where the prior distribution, $p(\boldsymbol{\theta})$, is a poor approximation of the actual target distribution.

The PMC sampler of *Turner and van Zandt* [2012] assumes that the sequence of $\boldsymbol{\epsilon}$ values is specified by the user. Practical experience suggests that a poor selection of $\boldsymbol{\epsilon} = \{\epsilon_1, \ldots, \epsilon_J\}$ can lead to very low acceptance rates or even premature convergence. *Sadegh and Vrugt* [2013] have therefore introduced an alternative variant of ABC-PMC with adaptive selection of $\epsilon_{j(j>1)}$. This method requires the user to specify only the initial kernel bandwidth, ϵ_1, and subsequent values of $\boldsymbol{\epsilon}$ are determined from the $\rho(\cdot)$ values of the N most recently accepted samples. This approach is not only more practical, but also enhances convergence speed to the posterior distribution. We therefore prescribe the sequence of $\boldsymbol{\epsilon}$ values in ABC-PMC using the outcome of several adaptive runs.

It is interesting to note that the PMC sampler has elements in common with genetic algorithms (GA) [*Higuchi*, 1997] in that a population of individuals is used to search the parameter space. The main difference between both approaches is that PMC is specifically designed for statistical inference of the marginal and joint parameter distributions, whereas GAs are specialized in optimization. Yet, it is not difficult to modify the PMC sampler so that it converges to a single "best" solution. Nonetheless, one should be particularly careful using common GA operators such as crossover and mutation. Such genetic operators can significantly improve the search capabilities of ABC-PMC in high dimensional search spaces, but can harm convergence properties.

To benchmark the efficiency of ABC-REJ and ABC-PMC, we start by fitting a relatively simple mixture of two Gaussian distributions, which has become a classical problem in the ABC literature [*Sisson et al.*, 2007, *Beaumont et al.*, 2009, *Toni et al.*, 2009, *Turner and*

Sederberg, 2012]

$$p(\theta) = \frac{1}{2}\mathcal{N}(0, \frac{1}{100}) + \frac{1}{2}\mathcal{N}(0, 1), \tag{4.5}$$

where $\mathcal{N}(a, b)$ is a normal distribution with mean, a and standard deviation, b. The solid black line in Figure 4.2 plots the actual target distribution.

We now test the efficiency of ABC-REJ and ABC-PMC using the following distance function,

$$\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}^*))\big) = \begin{cases} |\overline{\mathbf{Y}}| & \text{with probability of } 50\% \\ |y_1| & \text{with probability of } 50\% \end{cases}$$

where $\mathbf{Y} = \{y_1, \ldots, y_{100}\}$, $y_i \sim \mathcal{N}(\theta, 1)$, and the operator $|\cdot|$ signifies the modulus (absolute value). In keeping with the statistical literature, we assume a uniform prior distribution, $p(\theta) \sim \mathcal{U}[-10, 10]$, with $\epsilon = 0.025$ (ABC-REJ) and $\boldsymbol{\epsilon} = \{1, 0.75, 0.5, 0.25, 0.1, 0.05, 0.025\}$ (ABC-PMC). Figure 4.2 plots histograms of the ABC-REJ (A: left plot) and ABC-PMC (B: middle plot) derived posterior distribution of θ using $N = 1,000$ samples.

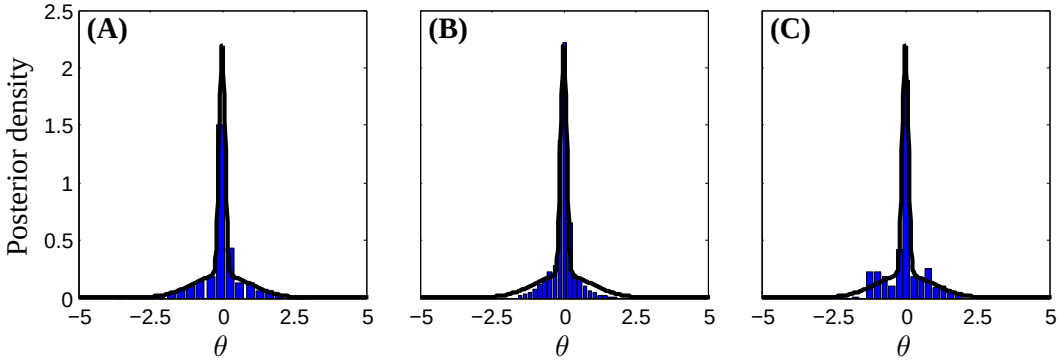


Figure 4.2: One-dimensional mixture distribution (solid black line) and histogram of the posterior samples derived from (A) Rejection sampling (ABC-REJ) , (B) Population Monte Carlo sampling (ABC-PMC), and (C) MCMC simulation with $\text{DREAM}_{(\text{ABC})}$.

The marginal distribution derived with ABC-REJ and ABC-PMC are in good agreement with the known target distribution (black line). Table 4.1 summarizes the performance of

the ABC-REJ and ABC-PMC sampler. The rejection sampler (ABC-REJ) requires about $386,742$ function evaluations to find $N = 1,000$ behavioral solutions. This corresponds to an acceptance rate (AR, %) of approximate 0.26%, which can be considered highly inefficient. The ABC-PMC sampler on the other hand requires fewer function evaluations (170,848) to explore the target distribution, with an acceptance rate of about 0.59%. Note however that ABC-PMC underestimates the sampling density of points in the tails of the posterior suggesting that the algorithm has not been able to fully explore $p(\theta)$. The results of DREAM$_{(\text{ABC})}$ that are listed at the bottom of Table 4.1 will be discussed in section 4.4.1 of this chapter.

Table 4.1: Case study I: 1-dimensional toy example. We list the (final) epsilon value, acceptance rate, AR (%) and number of function evaluations needed to sample the target distribution.

	ϵ	AR(%)	Func. Eval.
ABC-REJ	0.025	0.259	386,742
ABC-PMC	0.025	0.585	170,848
DREAM$_{(\text{ABC})}$	0.025	1.708	50,000

4.3 Markov Chain Monte Carlo Simulation

The adaptive capabilities of the ABC-PMC sampler offer significant computational advantages over ABC-REJ. However, further methodological improvements are warranted to enable inference of complex simulation models involving high dimensional parameter spaces. The use of a boxcar fitness kernel (zero probability everywhere except for a small interval where it is a constant) is theoretically convenient, but makes it very difficult for any sampling algorithm to determine the preferred search direction. All rejected simulations receive a similar score, irrespective of whether their $\rho(\cdot)$ values are in close proximity of the threshold, ϵ or far removed. This is certainly not desirable and unnecessarily complicates posterior exploration. Furthermore, ABC-REJ and ABC-PMC update all entries of the parameter vector simultaneously. This is equivalent to a crossover of 100%, and adequate for low-

dimensional problems involving just a handful of parameters, but not necessarily efficient in high dimensional search spaces. For such problems, conditional (or subspace) sampling has desirable properties and can enhance, sometimes dramatically, the speed of convergence. The method we propose herein is more commonly known as Metropolis-within-Gibbs, and samples individual dimensions (or groups of parameters) in turn.

4.3.1 Continuous fitness kernel

A boxcar kernel has the disadvantage that all samples with $\rho(\cdot)$ value larger than ϵ are considered equal and discarded. This is the basis of rejection sampling, and can be very inefficient particularly if the prior sampling distribution is poorly chosen. To solve this problem, *Turner and Sederberg* [2012] recently introduced an alternative ABC sampling approach using MCMC simulation with a continuous fitness kernel. This method is based on the concept of noisy-ABC [*Beaumont et al.*, 2002, *Blum and François*, 2010] and perturbs the model simulation with a random error, $\boldsymbol{\xi}$

$$\mathbf{Y} \leftarrow \text{Model}(\boldsymbol{\theta}^*) + \boldsymbol{\xi} \tag{4.6}$$

If we assume that $\boldsymbol{\xi}$ follows a multivariate normal distribution, $\mathcal{N}(\mathbf{0}_n, \alpha)$, then we can evaluate the probability density, $p\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}^*))\big)$, of $\boldsymbol{\theta}^*$ using

$$p(\boldsymbol{\theta}^*|\alpha) = \frac{1}{\alpha\sqrt{2\pi}} \exp\left(-\frac{1}{2}\alpha^{-2}\Big(\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}^*))\big)\Big)^2\right), \tag{4.7}$$

where α is an algorithmic (or free) parameter that defines the width of the kernel. As a result, the approximation in Equation (4.3) becomes [*Turner and Sederberg*, 2012]

$$p(\boldsymbol{\theta}|\tilde{\mathbf{Y}}) \propto \int_{\boldsymbol{\mathcal{Y}}} p(\boldsymbol{\theta})\text{Model}(\boldsymbol{y}|\boldsymbol{\theta})p\Big(\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}^*))\big)\Big)d\boldsymbol{y} \tag{4.8}$$

This approach has recently been coined kernel-based ABC (KABC) in the statistical literature, and opens up an arsenal of advanced Monte Carlo based sampling methods to explore the posterior distribution. Three preliminary case studies of different complexity (not shown herein) demonstrate that KABC with DREAM is at least $3-1,000$ times more efficient than ABC-PMC. Unfortunately, KABC can run into a fatal problem. The MCMC algorithm may produce a nice bell shaped posterior distribution but with simulated summary statistics that are far removed from their observed values. For example, let's assume an extreme case in which the model is unable to fit any of the observed summary statistics within the required tolerance ϵ. Sampling would still produce a limiting distribution, but with probability densities of Equation (4.7) that are practically zero. These results are undesirable, and have nothing to do with the actual MCMC method used, replacing this with another sampling approach would give similar findings. The culprit is the continuous kernel of Equation (4.7), which does not a-priori bound the feasible space of the posterior solutions. Changing Equation (4.7) to a boxcar function with $p(\cdot) = 1$ in the interval $[-\epsilon, \epsilon]$ around the measured summary statistic(s), and exponential decline of the density outside this interval runs into the same problems and is thus also futile.

We here propose an alternative method for fitness assignment that helps convergence of a MCMC simulator to the posterior distribution. We define the fitness of $\boldsymbol{\theta}^*$ as follows

$$f(\boldsymbol{\theta}^*, \phi) = \phi - \rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}^*))\big), \tag{4.9}$$

where $\phi > 0$ is a coefficient that bounds the fitness values between $(-\infty, \phi]$. The smaller the distance of the simulated summary metrics to their observed values, the higher the fitness. By setting $\phi = \epsilon$, then $f(\boldsymbol{\theta}^*) \in [0, \epsilon]$ is a necessary condition for a sample, $\boldsymbol{\theta}^*$, to be called a posterior solution, otherwise the sample is non behavioral and can be discarded. This condition is easily verified a-posteriori from the sampled fitness values of the Markov chains.

We are now left with a definition of the selection rule to help determine whether to accept trial moves or not. The original scheme proposed by *Metropolis et al.* [1953] was constructed using the condition of detailed balance. If $p(u)$ $(p(i))$ denotes the probability to find the system in state u (i) and $q(u \to i)\,(q(i \to u))$ is the conditional probability to perform a trial move from u to i (i to u), then the probability $p_{\text{acc}}(u \to i)$ to accept the trial move from u to i is related to $p_{\text{acc}}(i \to u)$ according to:

$$p(u)q(u \to i)p_{\text{acc}}(u \to i) = p(i)q(i \to u)p_{\text{acc}}(i \to u) \tag{4.10}$$

If we assume a symmetric jumping distribution, that is $q(u \to i) = q(i \to u)$, then it follows that

$$\frac{p_{\text{acc}}(u \to i)}{p_{\text{acc}}(i \to u)} = \frac{p(i)}{p(u)} \tag{4.11}$$

This Equation does not yet determine the acceptance probability. *Metropolis et al.* [1953] made the following choice:

$$p_{\text{acc}}(u \to i) = \min\left[1, \frac{p(i)}{p(u)}\right], \tag{4.12}$$

to determine whether to accept a trial move or not. This selection rule has become the basic building block of many existing MCMC algorithms. *Hastings* [1970] extended Equation (4.12) to non-symmetrical jumping distributions in which $q(u \to i) \neq q(i \to u)$.

Unfortunately, Equation (4.9) is not a proper probability density function, and hence application of (4.12) will lead to a spurious approximation of the posterior distribution. The same problem arises with other fitness functions derived from Equation (4.7), for example, the Nash-Sutcliffe efficiency, pseudolikelihoods and a composite of multiple objective functions.

We therefore calculate the acceptance probability using

$$p_{\text{acc}}(u \rightarrow i) = \max\Big(\mathrm{I}\big(f(i) \geq f(u)\big), \mathrm{I}\big(f(i) \geq (\phi - \epsilon)\big)\Big) \tag{4.13}$$

where $\mathrm{I}(\cdot)$ is an indicator function, and the operator $\max\big(\mathrm{I}(a), \mathrm{I}(b)\big)$ returns one if a and/or b is true, and zero otherwise. Thus, $p_{\text{acc}}(u \rightarrow i) = 1$ (we accept) if the fitness of the proposal i is higher than or equal to that of the current state of the chain, u. On the contrary, if the fitness of i is smaller than that of u then $p_{\text{acc}}(u \rightarrow i) = 0$ and the proposal is rejected, unless $f(i) \geq 0$, then we still accept.

The binary acceptance probability of Equation (4.13) differs fundamentally from a regular MCMC selection rule, but has important practical advantages that promulgate converge to the correct limiting distribution. Initially, when the samples are rather inferior, Equation (4.13) enforces the MCMC algorithm to act as an optimizer and only accept proposals with a higher fitness and thus smaller distance to the observed values of the summary statistics. The transitions of the Markov chain during this time are irreversible with a zero acceptance probability of the previous state (backward jump). This changes the moment a candidate point has been sampled whose fitness, $f(\cdot) \geq 0$. From this point forward, the acceptance probability of Equation (4.13) leads to a reversible Markov chain and the successive samples can be used to approximate the posterior target distribution.

An important limitation of Equation (4.13) is that it cannot incorporate non-symmetric jump distributions, such as the snooker updater used in DREAM$_{\text{(ZS)}}$ and MT-DREAM$_{\text{(ZS)}}$ [*Laloy and Vrugt*, 2012]. This would require a Hastings-type correction, but cannot be readily incorporated within the current framework.

4.3.2 Pseudo-code of DREAM$_{(\text{ABC})}$

We can solve for the posterior distribution of Equation (4.9) using MCMC simulation with DREAM [*Vrugt et al.*, 2008a, 2009a]. This method uses subspace sampling to increase search efficiency and overcome some of the main limitations of ABC-REJ and ABC-PMC. In DREAM, K ($K > 2$) different Markov chains are run simultaneously in parallel, and multivariate proposals are generated on the fly from the collection of chains, $\boldsymbol{\Theta}^{t-1}$ (matrix of $K \times d$ with each chain state as row vector), using differential evolution [*Storn and Price*, 1997, *Price et al.*, 2005]. If A is a subset of δ-dimensional space of the original parameter space, $\mathbb{R}^{\delta} \subseteq \mathbb{R}^{d}$, then a jump in the kth chain, $k = \{1, \ldots, K\}$ at iteration $t = \{2, \ldots, T\}$ is calculated using

$$\begin{aligned} \boldsymbol{\Delta}^{*}_{k,A} &= (\mathbf{1}_{\delta} + \boldsymbol{\lambda})\gamma(\delta)\left[\sum_{j=1}^{\tau} \boldsymbol{\theta}^{t-1}_{\mathbf{g}_j,A} - \sum_{j=1}^{\tau} \boldsymbol{\theta}^{t-1}_{\mathbf{r}_j,A}\right] + \boldsymbol{\zeta} \\ \boldsymbol{\Delta}^{*}_{k,\neq A} &= 0, \end{aligned} \tag{4.14}$$

where $\gamma = 2.38/\sqrt{2\tau D}$ is the jump rate, τ denotes the number of chain pairs used to generate the jump, and $\mathbf{g}$ and $\mathbf{r}$ are τ-vectors with integer values drawn without replacement from $\{1, \ldots, k-1, k+1, \ldots, K\}$. The values of $\boldsymbol{\lambda}$ and $\boldsymbol{\zeta}$ are sampled independently from $\mathcal{U}_{\delta}(-c, c)$ and $\mathcal{N}_{\delta}(0, c^{*})$ with, typically, $c = 0.1$ and c^{*} small compared to the width of the target distribution, $c^{*} = 10^{-12}$ say.

The candidate point of chain k at iteration t then becomes

$$\boldsymbol{\theta}^{*}_{k} = \boldsymbol{\theta}^{t-1}_{k} + \boldsymbol{\Delta}^{*}_{k}, \tag{4.15}$$

and the Metropolis ratio is used to determine whether to accept this proposal or not. The DREAM algorithm solves an important practical problem in MCMC simulation, namely that of choosing an appropriate scale and orientation of the proposal distribution. Section

4.3.3 will detail the procedure for selecting the subset A of dimensions of the parameter space that will be updated each time a proposal is created.

We now proceed with a pseudo-code of DREAM$_{(\mathrm{ABC})}$.

Algorithm 8 DREAM$_{(\text{ABC})}$-Markov chain Monte Carlo sampler

At iteration $t = 1$,
for $k = 1, \ldots, K$ **do**
  Sample $\boldsymbol{\theta}_k^1$ from the prior, $\boldsymbol{\theta}_k^1 \sim p(\boldsymbol{\theta})$
  Simulate data $\mathbf{Y}$ using $\boldsymbol{\theta}_k^1$, $\mathbf{Y} \sim \text{Model}(\boldsymbol{\theta}_k^1)$
  Calculate the fitness, $f(\boldsymbol{\theta}_k^1) = \epsilon - \rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}_k^1))\big)$
end for
At iteration $t > 1$,
for $t = 2, \ldots, T$ **do**
  for $k = 1, \ldots, K$ **do**
    Determine subset A, the dimensions of the parameter space to be updated.
    Calculate the jump vector, $\boldsymbol{\Delta}_k^*$ using Equation (4.14)
    Compute the proposal, $\boldsymbol{\theta}_k^* = \boldsymbol{\theta}_k^{t-1} + \boldsymbol{\Delta}_k^*$
    Simulate data $\mathbf{Y}$ using $\boldsymbol{\theta}_k^*$, $\mathbf{Y} \sim \text{Model}(\boldsymbol{\theta}_k^*)$
    Calculate the fitness, $f(\boldsymbol{\theta}_k^*) = \epsilon - \rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}_k^*))\big)$
    Calculate the acceptance probability using Equation (4.13),

$$p_{\text{acc}}(\boldsymbol{\theta}_k^*) = \max\Big(\text{I}\big(f(\boldsymbol{\theta}_k^*) \geq f(\boldsymbol{\theta}_k^{t-1})\big), \text{I}\big(f(\boldsymbol{\theta}_k^*) \geq 0\big)\Big)$$

    If $p_{\text{acc}}(\boldsymbol{\theta}_k^*) = 1$, set $\boldsymbol{\theta}_k^t = \boldsymbol{\theta}_k^*$ otherwise remain at current state, $\boldsymbol{\theta}_k^t = \boldsymbol{\theta}_k^{t-1}$
  end for
  Compute $\hat{R}$-statistic for each entry of $\boldsymbol{\theta}$ using last 50% of samples in each chain.
  If $\hat{R}_j \leq 1.2$ for all $j = \{1, \ldots, d\}$ stop, otherwise continue.
end for

In summary, DREAM$_{(\text{ABC})}$, runs K different Markov chains in parallel. Multivariate proposals in each chain $k = \{1, \ldots, K\}$ are generated by taking a fixed multiple of the difference of two or more randomly chosen members (chains) of $\boldsymbol{\Theta}$ (without replacement). By accepting each jump with binary probability of Equation (4.13) a Markov chain is obtained, the stationary or limiting distribution of which is the posterior distribution. Because the joint pdf of the K chains factorizes to $\pi(\boldsymbol{\theta}_1) \times \ldots \times \pi(\boldsymbol{\theta}_K)$, the states of the individual chains are independent at any iteration after DREAM$_{(\text{ABC})}$ has become independent of its initial value. After this burn-in period, the convergence of DREAM$_{(\text{ABC})}$ can thus be monitored with the $\hat{R}$-statistic of *Gelman and Rubin* [1992].

The jump distribution in Equation (4.14) of DREAM$_{(\text{ABC})}$ is easily implemented in ABC-PMC to help generate trial moves. This could further improve the scale and orientation of

the proposals, but comes at an increased computational cost. The conditional probability to move from $\boldsymbol{\theta}_{j-1}^{u}$ to $\boldsymbol{\theta}_{j}^{i}$ or $q_d(\boldsymbol{\theta}_{j-1}^{u} \rightarrow \boldsymbol{\theta}_{j}^{i})$ in Equation (4.4) of ABC-PMC is easy to calculate for a (multi)normal proposal distribution, but requires significantly more CPU-resources if the jumping kernel of DREAM$_{\text{(ABC)}}$ is used. Let's assume, for instance, that $\tau = 1$ and $\boldsymbol{\lambda} = \mathbf{0}_{\delta}$ in Equation (4.14). The probability to transition in ABC-PMC from the current state, $\boldsymbol{\theta}_{j-1}^{u}$, to the proposal, $\boldsymbol{\theta}_{j}^{i}$, is then equivalent to

$$q(\boldsymbol{\theta}_{j}^{i}|\boldsymbol{\theta}_{j-1}^{u}) = \sum_{m=1}^{N}\sum_{o=1}^{N}\psi(\boldsymbol{\theta}_{j-1}^{u} + \gamma(\boldsymbol{\theta}_{j-1}^{m} - \boldsymbol{\theta}_{j-1}^{o})|\boldsymbol{\Sigma}_d);\ i \neq o \neq m \tag{4.16}$$

where ψ denotes the normal density with covariance matrix $\boldsymbol{\Sigma}_d = (c^*)^2 I_d$. This Equation is of computational complexity $\mathcal{O}(N^2)$ and becomes particularly CPU-intensive for large N and/or if more than one pair of chains ($\tau > 1$) is used to create proposals. More fundamentally, the lack of subspace sampling (see next section) in ABC-PMC deteriorates search efficiency in high-dimensional spaces (shown later). We therefore do not consider this alternative jumping distribution in ABC-PMC.

4.3.3 Randomized subspace sampling

Subspace sampling is implemented in DREAM$_{\text{(ABC)}}$ by only updating randomly selected dimensions of $\boldsymbol{\theta}_{k}^{t-1}$ each time a proposal is generated. Following the default of the DREAM suite of algorithms [*Vrugt et al.*, 2008a, 2009a, 2011, *Laloy and Vrugt*, 2012] we use a geometric series of n_{CR} different crossover values and store this in a vector, CR $= \{\frac{1}{n_{\text{CR}}}, \frac{2}{n_{\text{CR}}}, \ldots, 1\}$. The prior probability of each crossover value is assumed equal and defines a vector $\mathbf{p}$ with n_{CR} copies of $\frac{1}{n_{\text{CR}}}$. We create the set A with selected dimensions to be updated as follows

Algorithm 9 Subspace sampling

for $k = 1, \ldots, K$ **do**
  Define A to be an empty set, $A = \emptyset$
  Sample P from the discrete multinomial distribution, $P \sim \mathfrak{F}(\mathrm{CR}|\mathbf{p})$
  Draw d labels from a multivariate uniform distribution, $\mathbf{Z} \sim \mathcal{U}_d[0, 1]$
  for $j = 1, \ldots, d$ **do**
    if $\mathbf{Z}_j > (1 - P)$ **then**
      Add dimension j to A
    end if
  end for
  if $A = \emptyset$ **then**
    Choose one index from {1,. . . ,d} and add to A
  end if
end for

The number of dimensions stored in A ranges between 1 and d and depends on the sampled value of the crossover. This relatively simple randomized selection strategy enables single-site Metropolis sampling (one dimension at a time), Metropolis-within-Gibbs (one or a group of dimensions) and regular Metropolis sampling (all dimensions). To enhance search efficiency, the probability of each n_{CR} crossover values is tuned adaptively during burn-in by maximizing the normalized Euclidean distance between successive states of the K chains [*Vrugt et al.*, 2009a]. The only algorithmic parameter that needs to be defined by the user is n_{CR}, the number of crossover values used. We use the standard settings of DREAM and use $n_{\mathrm{CR}} = 3$ in all the calculations reported herein. This concludes the algorithmic description of DREAM$_{(\mathrm{ABC})}$.

4.4 Numerical experiments

The next section compares the efficiency of ABC-REJ, ABC-PMC and DREAM$_{(\mathrm{ABC})}$ for two synthetic and one real-world experiment. These case studies cover a diverse set of problem features, including high-dimensionality, nonlinearity, non-convexity, and numerous local optima. Perhaps not surprisingly, our trials with ABC-REJ show that rejection sampling is

highly inefficient when confronted with multidimensional parameter spaces, unless the prior sampling distribution closely mimics the target distribution of interest. In practice, this is an unreasonable expectation and we therefore discard the ABC-REJ algorithm after the first case study and focus our attention on the results of ABC-PMC and DREAM$_{(\text{ABC})}$.

In all our calculations with ABC-PMC we create $N = 1,000$ samples at each iteration, $j = \{1, \ldots, J\}$ using values of $\boldsymbol{\epsilon}$ that are listed in each case study and have been determined through trial-and-error (e.g. *Sadegh and Vrugt* [2013]). In DREAM$_{(\text{ABC})}$ we need to define the number of chains, K and the total number of function evaluations, $M = K \cdot T$. Their values are listed in each individual case study. For all the other algorithmic variables in DREAM$_{(\text{ABC})}$ we use standard settings recommended in *Vrugt et al.* [2009a].

4.4.1 Synthetic benchmark experiments: Gaussian mixture model

We now benchmark the performance of DREAM$_{(\text{ABC})}$ by application to the Gaussian mixture model in Equation (4.5). We set $\epsilon = 0.025$ and run $K = 10$ different chains using a total of $M = 50,000$ function evaluations. Figure 4.2c displays the marginal distribution of the posterior samples using a burn-in of 50%. It is evident that the adaptive capabilities of DREAM$_{(\text{ABC})}$ enables it to track the target distribution. The density of samples in the tails has somewhat improved considerably compared to ABC-PMC. The burn-in required for multi-chain methods such as DREAM$_{(\text{ABC})}$ is relatively costly for this one-dimensional problem, and hence it is not surprising that rejection sampling provides a nicer sample of the mixture distribution. The DREAM$_{(\text{ABC})}$ approximation of the target is easily enhanced by creating more samples.

Table 4.1 lists the acceptance rate (AR, %), number of function evaluations, and ϵ value used with ABC-REJ, ABC-PMC and DREAM$_{(\text{ABC})}$. It is evident that DREAM$_{(\text{ABC})}$ is most efficient in sampling the target distribution. The acceptance rate of DREAM$_{(\text{ABC})}$ is

between 3 to 6 times higher than that ABC-PMC and ABC-REJ, respectively.

4.4.2 Synthetic benchmark experiments: 20-dimensional bivariate distribution

A more demanding test of the ABC-PMC and DREAM$_{(\text{ABC})}$ algorithms can be devised by using a multi-dimensional target distribution. We consider a set of ten bivariate normal distributions

$$h_i(\mu_{i,1}, \mu_{i,2}) \sim \mathcal{N}_2 \left(\begin{bmatrix} \mu_{i,1} \\ \mu_{i,2} \end{bmatrix}, \begin{bmatrix} 0.01^2 & 0 \\ 0 & 0.01^2 \end{bmatrix} \right), \tag{4.17}$$

with unknown mean of the ith component, $\boldsymbol{\mu}_i = \{\mu_{i,1}, \mu_{i,2}\}$ and fixed covariance matrix. We now generate $n = 20$ observations by sampling the mean of each bivariate distribution from $\mathcal{U}_2[0, 10]$. The "observed" data are plotted in Figure 4.3 using the red cross symbols. Each of the bivariate means is now subject to inference with ABC, which results in a $d = 20$ dimensional parameter estimation problem. The simulated data, $\mathbf{Y}$ are created by evaluating Equation (4.17) fifty different times for each proposal, $\boldsymbol{\theta}^* = [\boldsymbol{\mu}_1^*, \ldots, \boldsymbol{\mu}_{10}^*]$. The mean of the fifty samples from each bivariate distribution is stored in $\mathbf{Y}$ (10 by 2 matrix) and compared to the observed data using the following distance function [*Turner and Sederberg*, 2012]

$$\rho\big(\tilde{\mathbf{Y}}, \mathbf{Y}(\boldsymbol{\theta}^*)\big) = \sqrt{\frac{1}{20} \sum_{i=1}^{10} \sum_{j=1}^{2} \big(\tilde{\mathbf{Y}}_{(i,j)} - \mathbf{Y}_{(i,j)}(\boldsymbol{\theta}^*)\big)^2} \tag{4.18}$$

We assume a noninformative (uniform) prior, $\boldsymbol{\theta} \sim \mathcal{U}_{20}[0, 10]$ and set

$$\boldsymbol{\epsilon} = \{3, 2.5, 2.1, 1.8, 1.6, 1.3, 1.1, 0.9, 0.8, 0.7, 0.6\}$$

Table 4.2: Case study II: 20-dimensional bivariate Gaussian distribution. We list the final epsilon value, acceptance rate, AR (%) and number of function evaluations needed to sample a stationary distribution.

	ϵ	AR(%)	Func. Eval.
ABC-PMC	0.600	0.019	5,143,989
DREAM$_{(\text{ABC})}$	0.025	19.717	200,000

(ABC-PMC) and $\epsilon = 0.025$, $K = 15$, and $M = 200,000$ (DREAM$_{(\text{ABC})}$). The results of the analysis are presented in Table 4.2 and Figure 4.3.

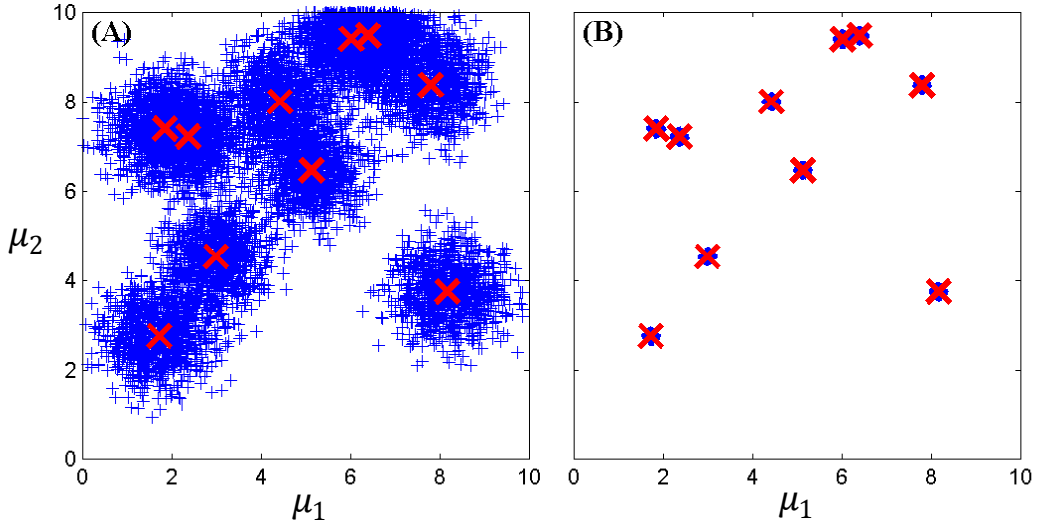


Figure 4.3: Two-dimensional scatter plots of the posterior samples, "+" generated with (A) Population Monte Carlo sampling, and (B) MCMC simulation with DREAM$_{(\text{ABC})}$. The true values of μ_1 and μ_2 are indicated with a cross, "×". The ABC-PMC samples are very dispersed and significantly over estimate the actual width of the target distribution.

Figure 4.3 plots the posterior samples (plus symbol) derived from (a) ABC-PMC and (b) DREAM$_{(\text{ABC})}$. The sampled solutions cluster around the observed means (red cross) of the bivariate normal distributions. The size of the posterior uncertainty differ markedly between both algorithms. The posterior samples of DREAM$_{(\text{ABC})}$ are in excellent agreement with the bivariate target distributions. The samples group tightly around their observed counterparts, and their structure is in excellent agreement with the covariance matrix of the target distribution. The ABC-PMC samples, on the contrary, exhibit too much scatter.

This finding is not surprising. The ABC-PMC sampler terminated its search prematurely with values of $\rho(\cdot)$ between 0.5 and 0.6. This threshold is much larger than the value of $\epsilon = 0.025$ required to converge to the target distribution. Subspace sampling would significantly enhance the results of ABC-PMC (not shown), but such modifications could affect the theoretical convergence properties.

The DREAM$_{(\text{ABC})}$ algorithm not only better recovers the actual target distribution, but its sampling efficiency is also superior. To illustrate this in more detail, consider Table 4.2 that lists the acceptance rate (AR ,%), number of function evaluations, and ϵ value of ABC-PMC and DREAM$_{(\text{ABC})}$. The acceptance rate of DREAM$_{(\text{ABC})}$ of about 19.72% is more than $1,000$ times higher than that of ABC-PMC (0.019%). This marks a three-order of magnitude improvement in search efficiency, which in large part is due to the ability of DREAM$_{(\text{ABC})}$ to sample one or groups of variables in turn. This conditional sampling is necessary to traverse multi-dimensional parameter spaces in pursuit of the posterior distribution.

The present case study clearly illustrates the advantages of DREAM$_{(\text{ABC})}$ when confronted with multi-dimensional parameter spaces. The algorithm requires about $M = 200,000$ function evaluations to successfully recover the 20-d target distribution. To illustrate this in more detail, consider Figure 4.4 which displays trace plots of the $\hat{R}$-statistic of *Gelman and Rubin* [1992] using the last 50% of the samples stored in each of the K chains. This convergence diagnostic compares for each parameter the between- and within-variance of the chains. Because of asymptotic independence, the between-member variance and $\hat{R}$-diagnostic can be estimated consistently from a single DREAM$_{(\text{ABC})}$ trial. Values of $\hat{R}$ smaller than 1.2 indicate convergence to a limiting distribution. The DREAM$_{(\text{ABC})}$ algorithm needs about $40,000$ function evaluations to officially reach convergence and generate a sufficient sample of the posterior distribution. This is much less than the $M = 200,000$ function evaluations used in this study. Obviously, one should be careful to judge convergence of the sampled Markov chains based on a single diagnostic, yet visual inspection of the sampled trajectories confirms

an adequate mixing of the different chains and converge of DREAM$_{(\text{ABC})}$ after about $15,000$ function evaluations. This is substantially lower than the approximately $40,000$ function evaluations estimated with the $\hat{R}$-statistic, simply because the second half of the chain is used for monitoring convergence.

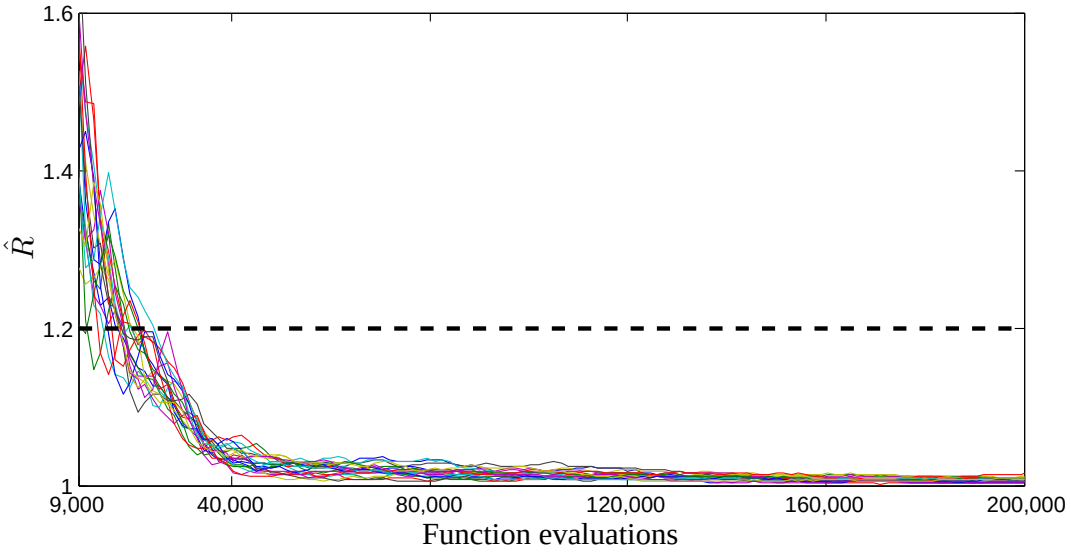


Figure 4.4: Trace plots of the $\hat{R}$-statistic of the sampled Markov chains with DREAM$_{(\text{ABC})}$ for the 20-dimensional bivariate normal distribution. Each parameter is coded with a different color. The dashed line denotes the default threshold used to diagnose convergence to a limiting distribution.

4.4.3 Hydrologic modeling: Sacramento Soil Moisture Accounting Model

A more realistic case study is now devised and used to illustrate the advantages DREAM$_{(\text{ABC})}$ can offer in real-world modeling problems. I consider simulation of the rainfall-runoff transformation using the SAC-SMA conceptual hydrologic model. This model has been developed by *Burnash et al.* [1973] and is used extensively by the National Weather Service for flood forecasting throughout the United States. The model has been described in detail in many previous publications, and we therefore summarize in Table 4.3 the fourteen parameters that

require calibration and their prior uncertainty ranges.

Table 4.3: Prior ranges of the SAC-SMA model parameters and their posterior mean values derived for the French Broad river basin data using the ABC-PMC and DREAM$_{(\text{ABC})}$ algorithms. We also summarize the results of a residual-based Gaussian likelihood function, DREAM$_{(\text{RBGL})}$.

Parameter	Range	Posterior Parameter Mean		
		ABC-PMC	DREAM$_{(\text{ABC})}$	DREAM$_{(\text{RBGL})}$
UZTWM	1-150	81.812	82.676	39.725
UZFWM	1-150	73.722	67.945	10.531
UZK	0.1-0.5	0.299	0.301	0.387
PCTIM	0-0.1	0.007	0.010	0.003
ADIMP	0-0.4	0.118	0.121	0.189
ZPERC	1-250	132.350	130.883	109.207
REXP	1-5	2.972	2.847	4.880
LZTWM	1-500	396.792	363.329	456.619
LZFSM	1-1000	282.049	246.326	146.584
LZFPM	1-1000	796.113	728.855	733.093
LZSK	0.01-0.25	0.145	0.142	0.124
LZPK	0.0001-0.025	0.006	0.006	0.009
PFREE	0-0.6	0.292	0.304	0.530
RRC	0-1	0.630	0.581	0.321

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 are used in the present study. In keeping with *Vrugt and Sadegh* [2013] we use the annual runoff coefficient, $S_1(\tilde{\mathbf{Y}})$ (-), the annual baseflow index, $S_2(\tilde{\mathbf{Y}})$ (-), and the flow duration curve, $S_3(\tilde{\mathbf{Y}})$ (day/mm) and $S_4(\tilde{\mathbf{Y}})$ (-), as summary metrics of the discharge data. A detailed description of each summary statistic is given by *Vrugt and Sadegh* [2013] and interested readers are referred to this publication for further details. This leaves us with $L = 4$ four summary statistics for three different hydrologic signatures.

We use the following composite distance function to quantify the distance between the observed and simulated summary statistics

$$\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}^*))\big) = \max\big(|S_i(\tilde{\mathbf{Y}}) - S_i(\mathbf{Y}(\boldsymbol{\theta}^*))|\big) \quad i = \{1, \ldots, L\}, \tag{4.19}$$

and help determine whether to accept $\boldsymbol{\theta}^*$ or not. Model simulations that simultaneously satisfy each of the four summary metrics within their tolerance thresholds are considered behavioral, and hence constitute samples from the posterior distribution. This composite formulation differs fundamentally from multi-criteria model calibration approaches in which the summary statistics (objective functions) are assumed non-commensurate and therefore treated independently from one another. This latter approach gives rise to a Pareto solution set (rather than posterior distribution) and quantifies the trade-offs in the fitting of the different metrics.

To maximize the search efficiency of ABC-PMC we use

$$\boldsymbol{\epsilon} = \{1, 0.3, 0.15, 0.1, 0.07, 0.06, 0.04, 0.025\}.$$

This sequence is determined from a preliminary run of ABC-PMC with adaptive selection of ϵ (see Appendix B of *Sadegh and Vrugt* [2013]). The DREAM$_{(\text{ABC})}$ sampler is executed using default values of the algorithmic parameters and $\epsilon = 0.025$, $K = 15$, and $M = 200,000$. Tables 4.3 and 4.4 and Figures 4.5, 4.6, and 4.7 summarize our main findings.

Table 4.4 compares the computational efficiency of ABC-PMC and DREAM$_{(\text{ABC})}$. For completeness, we also list the results of DREAM using a residual-based Gaussian likelihood function, hereafter referred to as DREAM$_{(\text{RBGL})}$. The DREAM$_{(\text{ABC})}$ algorithm has an acceptance rate (AR, %) of about 3.14% and requires $200,000$ SAC-SMA model evaluations to generate $40,000$ posterior samples. The ABC-PMC sampler, on the contrary, is far less efficient (AR = 0.046%) and needs about 2.2 million function evaluations to produce $1,000$ posterior samples. This constitutes a more than ten times difference in sampling efficiency, and favors the use of DREAM$_{(\text{ABC})}$ for diagnostic inference of complex and CPU-intensive models.

The acceptance rate of DREAM$_{(\text{RBGL})}$ of 4.54% is much lower than the theoretical (opti-

Table 4.4: Case study III: 14-dimensional SAC-SMA model calibration problem. We list the final epsilon value, acceptance rate, AR (%) and number of function evaluations needed to sample the posterior distribution. We also include the results of DREAM using a residual-based Gaussian likelihood function.

	ϵ	AR(%)	Func. Eval.
ABC-PMC	0.025	0.046	2,173,490
$\text{DREAM}_{(\text{ABC})}$	0.025	3.135	200,000
$\text{DREAM}_{(\text{RGBL})}$	N/A	4.543	200,000

mal) value of about 23.4% for the considered dimensionality of the target distribution. This finding is not surprising and can be explained by the non-ideal properties of the SAC-SMA response surface [*Duan et al.*, 1992], which, to a large extent, are inflicted by poor numerics [*Clark and Kavetski*, 2010, *Kavetski and Clark*, 2010, *Schoups et al.*, 2010b]. The use of an explicit, Euler-based, integration method introduces pits and local optima (amongst others) on the response surface, and their presence deteriorates the search efficiency of MCMC methods. An implicit, time-variable, integration method would give a smoother response surface but at the expense of an increase in CPU-time. This increase in computational cost, will however, be balanced by a decrease in the number of model evaluations needed for a MCMC algorithm to converge to a limiting distribution.

Figure 4.5 presents histograms of the marginal posterior distributions derived with ABC-PMC (top panel), $\text{DREAM}_{(\text{ABC})}$ (middle panel) and $\text{DREAM}_{(\text{RBGL})}$ (lower panel). We display the results of a representative set of six SAC-SMA parameters and plot, from left to right across each panel, the posterior distributions of PCTIM, ADIMP, LZFSM, LZFPM, and LZPK. The x-axis matches exactly the ranges of each parameter used in the (uniform) prior distribution.

The marginal distributions derived from both sampling methods are in good agreement, and exhibit similar functional shapes. This inspires confidence in the ability of $\text{DREAM}_{(\text{ABC})}$ to correctly sample the target distribution. Most histograms extent a large part of the prior distribution, which suggests that the parameters are poorly defined by calibration against the

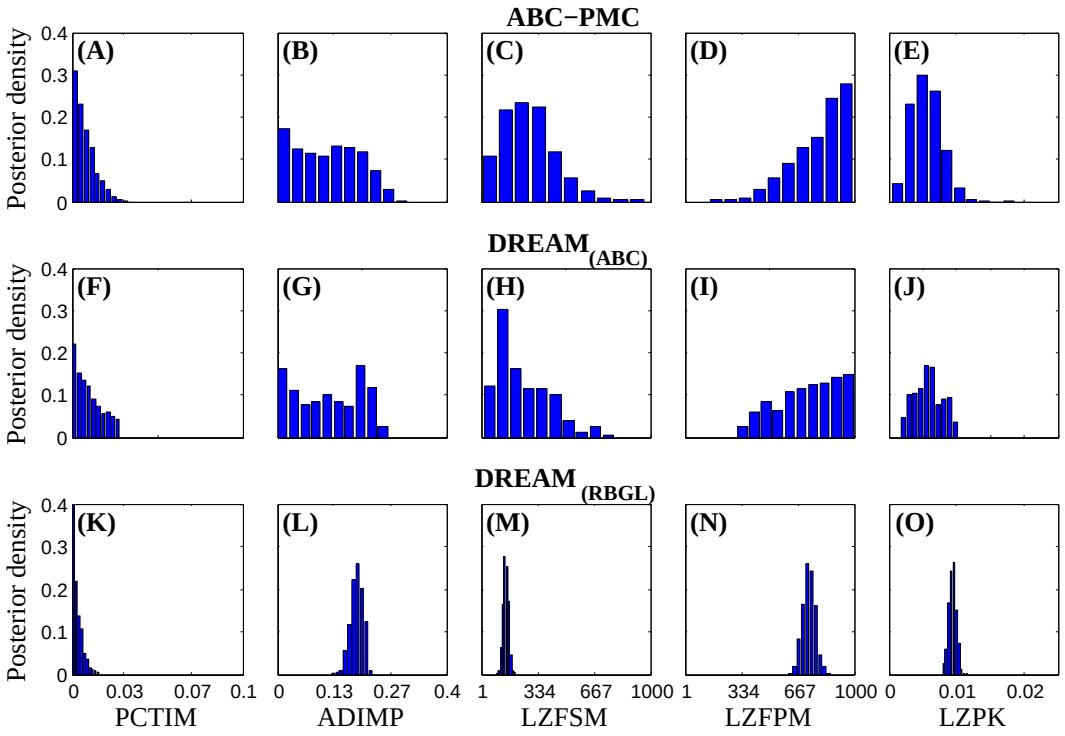


Figure 4.5: Marginal distributions of PCTIM, ADIMP, LZFSM, LZFPM and LZPK derived from the posterior samples created with ABC-PMC (upper panel) and DREAM$_{(ABC)}$ (middle panel). The histograms of the five SAC-SMA parameters occupy almost their entire prior distribution, which suggests that they are not particularly well identifiable by calibration against the observed annual baseflow index, annual runoff coefficient, and flow duration curve, respectively. The results of both ABC sampling methods are in good (visual) agreement, which inspires confidence in the ability of DREAM$_{(ABC)}$ to correctly sample the underlying target distribution. The bottom panel displays histograms of the SAC-SMA parameters derived using a classical residual-based (Gaussian) likelihood function. The SAC-SMA parameters are much better resolved, but these results cannot be justified given (amongst others) a lack of treatment of rainfall data errors.

four different summary statistics. This finding is perhaps not surprising. The four metrics used in this study are not sufficient, and extract only a portion of the information available in the discharge calibration data set. We will revisit this issue in the final paragraph of this section. Information theory can help to determine an approximate set of sufficient statistics, but this is beyond the scope of the present chapter.

We can further constrain the behavioral (posterior) parameter space by adding other signatures of catchment behavior to the current set of summary metrics. But, it is not particulary clear whether this would actually support the purpose of diagnostic model evaluation in which the (our) goal is not to just find the best possible fit of some model to some data set,

but rather to detect and pinpoint (epistemic) errors arising from inadequate or incomplete process representation. The chosen metrics appear relatively insensitive to rainfall data errors (not shown herein) and therefore exhibit useful diagnostic power. The bottom panel illuminates what happens to the SAC-SMA parameters if a least-squares type likelihood function is used for posterior inference. The parameters appear to be much better resolved by calibration against the observed discharge data but the remaining error residuals violate assumptions of homoscedasticity, normality and independence (not shown in detail). In part, this is due to a lack of treatment of rainfall data errors, whose probabilistic properties are difficult to accurately represent in a likelihood function. The generalized likelihood function of *Schoups and Vrugt* [2010a] provides ways to handle nontraditional residual distributions, nevertheless, this approach does not separate the contribution of individual error sources, and is therefore unable to provide insights into model malfunctioning. Note that the histograms of PCTIM, ADIMP, LZFSM, LZFPM, and LZPK are relatively tight and well described by a normal distribution, except for PCTIM which is hitting its lower bound.

To illustrate how the SAC-SMA posterior parameter uncertainty translates into modeled discharge uncertainty, please consider Figure 4.6 that presents time series plots of the 95% streamflow simulation uncertainty ranges (gray region) for a selected 500-day portion of the 3-year evaluation period derived from the posterior samples of ABC-PMC (top panel) and $\text{DREAM}_{(\text{ABC})}$ (middle panel). The observed discharge data are indicated with solid circles. Both time-series plots are in excellent agreement with simulated discharge dynamics that appear visually very similar and uncertainty ranges that envelop a large majority of the streamflow observations. Epistemic errors are not readily visible, yet this requires much further analysis possibly with the use of additional summary metrics. Previous results for this data set presented in *Vrugt and Sadegh* [2013], demonstrated an inability of the 7-parameter hmodel [*Schoups et al.*, 2010b] to simulate accurately the immediate response of the watershed to rainfall. This structural error can be resolved by model correction, a topic that will be studied in future publications. Note that the posterior mean RMSE derived

with DREAM$_{(\mathrm{ABC})}$ (0.72 mm/day) is somewhat lower than its counterpart from ABC-PMC (0.81 mm/day). This inconsistency conveys a difference in sampling density and posterior approximation.

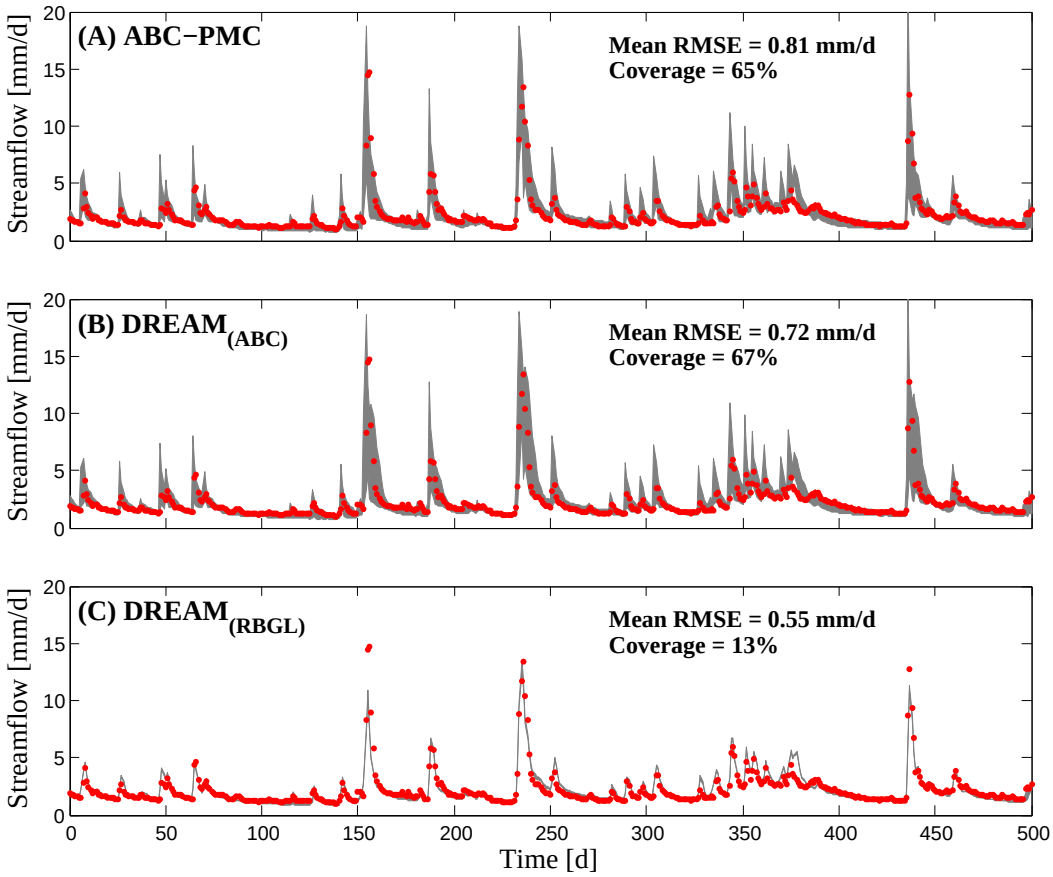


Figure 4.6: 95% posterior simulation uncertainty ranges (gray region) of the SAC-SMA model for a selected portion of the evaluation data set of the French Broad watershed. The top two panels displays the results of diagnostic model evaluation using (A) ABC-PMC, and (B) DREAM$_{(\mathrm{ABC})}$, whereas the bottom panel depicts the results of DREAM with a classical residual-based Gaussian likelihood function. The observed discharge values are indicated with the red dots. The SAC-SMA simulation intervals derived from ABC-PMC and DREAM$_{(\mathrm{ABC})}$ are very similar and encapsulate a large part of the discharge observations. The DREAM$_{(\mathrm{RBGL})}$ uncertainty ranges, on the other hand, exhibit a much lower coverage, but closer track the observed discharge data.

For completeness, the bottom panel of Figure 4.6 plots the 95% streamflow simulation uncertainty ranges derived from formal Bayes using a least-squares likelihood function. To enable a direct comparison with the results for diagnostic inference in the top two panels, we only consider the effect of parameter uncertainty on simulated discharge dynamics. The coverage

has decreased substantially to about 13%, which is hardly surprising given the relatively small width of the marginal distributions shown in Figure 4.5. The RMSE of the posterior mean SAC-SMA simulation (0.55 mm/day) is considerably lower than its counterparts derived from ABC-PMC and DREAM$_{(\mathrm{ABC})}$. This finding is not alarming but warrants some discussion. The four summary metrics used for diagnostic inference only extract partial information from the available discharge observations. This insufficiency makes it difficult to find a posterior model that "best" fits, in least-squares sense, the streamflow data, which is expected from a Gaussian likelihood function with homoscedastic measurement error. Also, the main purpose of diagnostic model evaluation with ABC is not that of model calibration, but rather to provide insights into model malfunctioning. Residual-based model calibration approaches provide little guidance on this issue which limits our ability to learn from the calibration data.

Table 4.5 presents summary variables (coverage, width, root mean square error, bias, and correlation coefficient) of the performance of the posterior mean SAC-SMA discharge simulation derived from the samples of ABC-PMC, DREAM$_{(\mathrm{ABC})}$ and DREAM$_{(\mathrm{RBGL})}$. We list results for the 5-year calibration and 3-year evaluation period. These statistics confirm our previous findings. The two different ABC sampling methods provide very similar results, and exhibit a better coverage of the discharge observations, larger width of the 95% simulation uncertainty ranges, and higher posterior mean RMSE than least-squares inference. The performance of the SAC-SMA model does not deteriorate during the evaluation period. In fact, the RMSE of the ABC derived posterior mean simulation substantially improves during the evaluation period, whereas this is not the case with least-squares fitting. This is a heartening prospect, and suggests (among others) that the chosen summary metrics at least partially represent the underlying signatures of watershed behavior.

To provide more insights into the convergence behavior of DREAM$_{(\mathrm{ABC})}$, Figure 4.7 plots the evolution of the $\hat{R}$-statistic of *Gelman and Rubin* [1992]. Each of the SAC-SMA parameters

is coded with a different color. About $160,000$ SAC-SMA model evaluations are required to converge to a limiting distribution. This marks a significant improvement in sampling efficiency over the ABC-PMC sampler which requires about 2.2 million function evaluations to create $1,000$ posterior samples.

We now turn our attention to the simulated values of the summary metrics. Figure 4.8 plots histograms of the posterior summary statistics derived with ABC-PMC (top panel) and DREAM$_{(\text{ABC})}$ (middle panel). The observed values of summary statistics are separately indicated in each plot with a red cross. The marginal distributions of summary metrics generally center around their measured values with the exception of the histogram of S_1 (annual runoff coefficient) that appears heavily skewed to the right. This points to a potential deficiency in the SAC-SMA model structure, yet this requires further analysis. For completeness, the bottom panel plots the posterior summary metric distributions derived from a residual-based likelihood function. This approach provides the closest fit to the observed streamflow data, but at the expense of summary metrics S_3 and S_4 (flow duration curve) that deviate considerably from their observed values. This finding highlights a profound difference in methodology between diagnostic model evaluation with ABC and residual-based inference. Of course, we could have used a different, and perhaps more reasonable, likelihood function for the error residuals [e.g. *Schoups and Vrugt* [2010a]]. This would affect some of our findings in Figure 4.8. Nevertheless, this is outside the scope of the present chapter, and we leave such comparison for future work.

Table 4.5: Performance of the SAC-SMA model for the calibration and evaluation data period of the French Broad river basin. We summarize the coverage (%) and average width (mm/d) of the 95% simulation intervals (due to parameter uncertainty), and the RMSE (mm/d), bias (%) and correlation coefficient, R (-) of the posterior mean SAC-SMA simulation.

	Coverage (%)		Width (mm/d)		RMSE (mm/d)		Bias (%)		R (-)	
	Calib.	Eval.	Calib.	Eval.	Calib.	Eval.	Calib.	Eval.	Calib.	Eval.
ABC-PMC	70.991	65.328	1.301	1.185	0.932	0.812	3.281	-3.721	0.863	0.832
$\text{DREAM}_{(\text{ABC})}$	71.100	66.515	1.408	1.272	0.831	0.722	3.731	-3.245	0.893	0.866
$\text{DREAM}_{(\text{RBGL})}$	18.829	12.500	0.155	0.141	0.539	0.545	2.767	-0.042	0.956	0.924

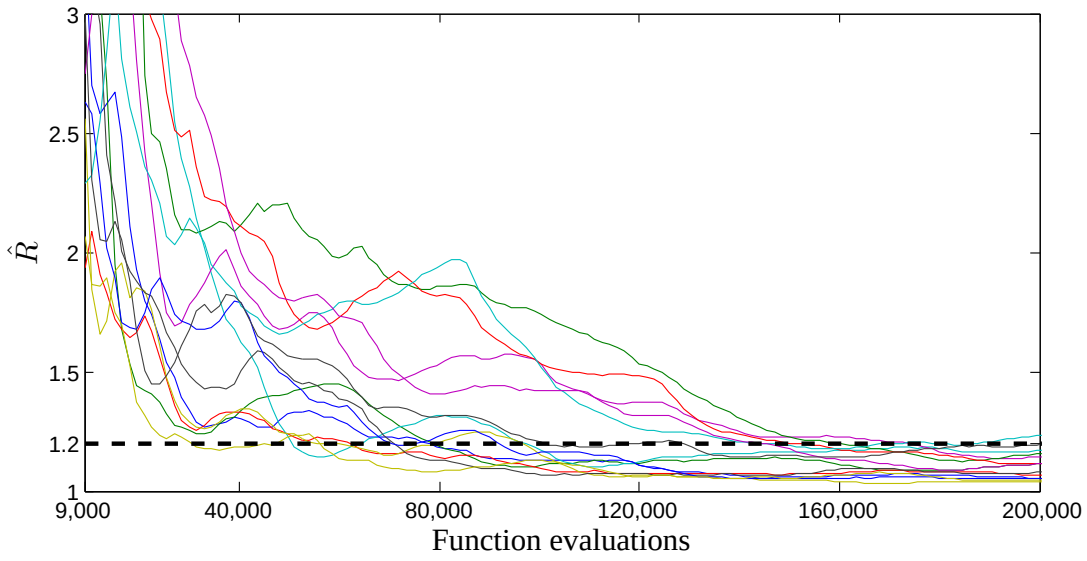


Figure 4.7: Evolution of the $\hat{R}$-statistic for the parameters of the SAC-SMA model using DREAM$_{(\text{ABC})}$ and discharge data from the French Broad watershed. Each of the parameters is coded with a different color. The dashed line denotes the default threshold used to diagnose convergence to a limiting distribution.

Finally, Figure 4.9 presents bivariate scatter plots of the posterior samples generated with DREAM$_{(\text{ABC})}$. We display the results for a representative set of all SAC-SMA parameter pairs including (A) PCTIM - LZPK, (B) LZSK - LZPK, (C) PCTIM - LZFPM, (D) PCTIM - LZSK, (E) LZPK - PFREE, and (F) ADIMP - LZPK. The axes in each of the six plots are in agreement with those used in the prior distribution. The bivariate posterior samples are confined to a densely-sampled rectangular (or square) space, and occupy a significant portion of the prior distribution. The white area immediately outside of the sampled space is made up of nonbehavioral solutions with fitness values smaller than zero (at least one summary metric is ϵ removed from its measured counterpart). The binary acceptance rule used in DREAM$_{(\text{ABC})}$ introduces a rather sharp demarcation of the behavioral solution space, nevertheless the sampled posterior distribution is in excellent agreement with its counterpart derived from ABC-PMC (see Figure 4.5). One could argue that that for this type of target distribution uniform random sampling should suffice. However, this method, which is at the heart of ABC-REJ, is highly inefficient in multi-dimensional parameter spaces. Many millions of function evaluations would be needed to provide a sufficient sample of the posterior

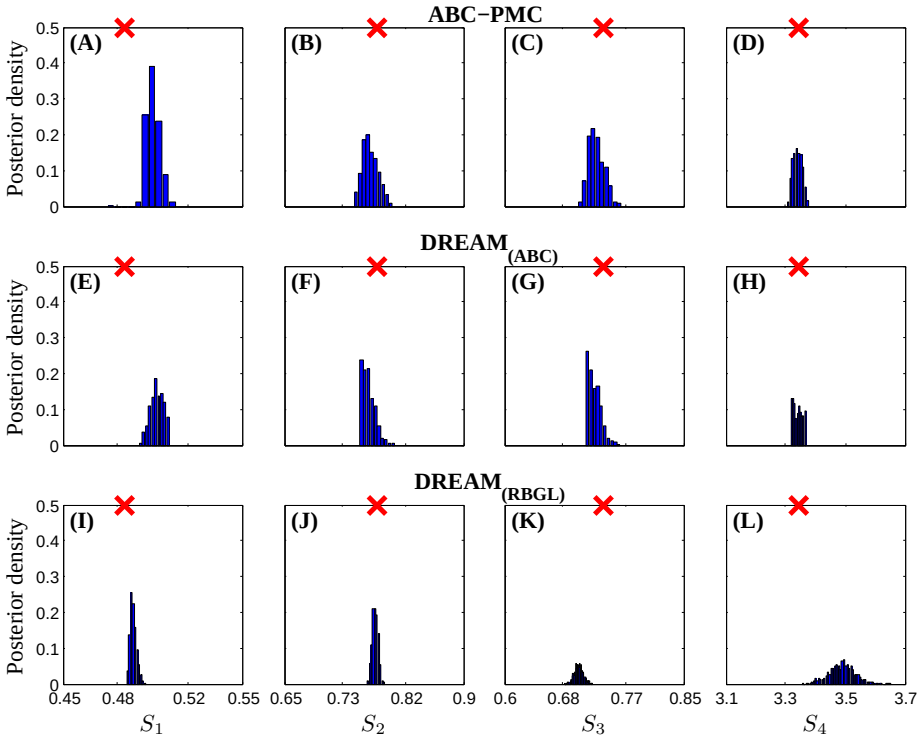


Figure 4.8: Histograms of the SAC-SMA derived summary statistics, S_1 (runoff), S_2 (baseflow), S_3 and S_4 (flow duration curve) of the posterior samples from ABC-PMC (top panel) and DREAM$_{(\mathrm{ABC})}$ (middle panel). The bottom panel displays the results of DREAM with a residual-based Gaussian likelihood function. The observed values of the summary metrics are separately indicated in each plot using the "$\times$" symbol. While $S_2 \rightarrow S_4$ center around their observed value ($\times$) for the ABC analysis, the marginal posterior distribution of S_1 is skewed to the right and does not encapsulate its measured value. This demonstrates that model is unable to simultaneously satisfy all the four different summary metrics used herein. This points to a structural deficiency in the SAC-SMA model structure, which will be investigated in more detail in subsequent work.

distribution. This is rather cumbersome, particularly if, as in the present case (not shown), the posterior samples exhibit parameter correlation.

The focus of the present chapter has been on improving ABC sampling efficiency to permit diagnostic inference of complex system models involving multi-dimensional parameter and summary metric spaces. Subsequent work can now focus on the intended purpose of diagnostic model evaluation and that is to help detect, diagnose, and resolve model structural deficiencies [*Vrugt and Sadegh*, 2013]. Commonly used model-data fusion approaches provide limited guidance on this important issue, in large part because of their aggregated treatment of input (forcing) data and epistemic errors. The use of summary statistics for

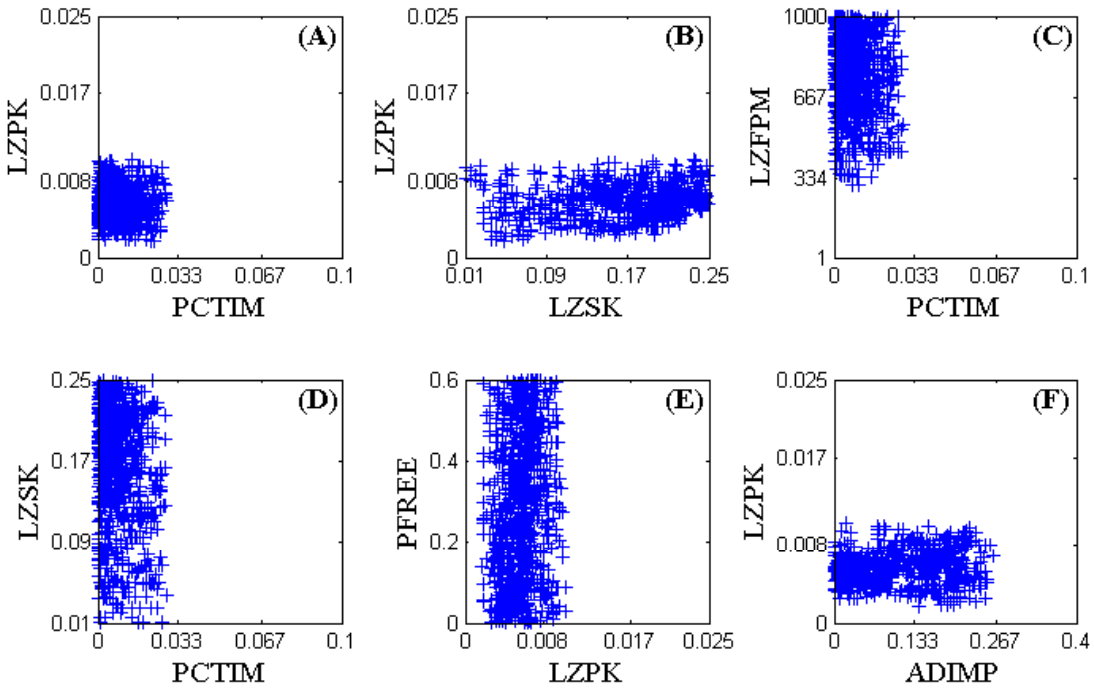


Figure 4.9: Two-dimensional scatter plots of the posterior samples, "+" generated with DREAM$_{(ABC)}$ for six different pairs of parameters of the SAC-SMA model. We restrict our attention to the (A) PCTIM - LZPK, (B) LZSK - LZPK, (C) PCTIM - LZFPM, (D) PCTIM - LZSK, (E) LZPK - PFREE, and (F) ADIMP - LZPK space, respectively. The bivariate posterior samples occupy a well-defined hypercube interior to the prior distribution.

statistical inference holds great promise, not only because signatures of system behavior are much less sensitive to, for instance, precipitation data errors than residual-based model fitting approaches, but also because the metrics can be devised in such a way that they relate directly to individual process descriptions and thus model components. This has important diagnostic advantages. Failure to fit one or more summary metrics can be directly addressed through correction of the responsible model component(s). This iterative process of inference, adjustment and refinement constitutes the basis of the scientific method. This also leaves the possibility to collect additional data to help validate new components of the model.

A recurrent issue with the application of diagnostic inference will be sufficiency of the summary metrics. Ideally, the summary metrics contain as much information as the original data

itself. Unfortunately, for most systems it will be rather difficult to find a set of sufficient summary statistics, unless each calibration data measurement is used as independent metric (e.g. *Sadegh and Vrugt* [2013]) but this defeats the purpose of diagnostic inference. Actually, it is not particularly clear whether sufficiency of the metrics is required to help detect and resolve epistemic errors. If deemed necessary, then one possible solution is to adapt formal Bayes and to use the summary metrics as an explicit prior. This type of approach has shown to significantly enhance the results of geophysical inversion [*Lochbühler et al.*, 2014c].

4.5 Summary and Conclusions

In Chapter 2, we have introduced approximate Bayesian computation (ABC) as vehicle for diagnostic model evaluation. Successful application of this methodology requires availability of an efficient sampling method that rapidly explores the space of behavioral models. Commonly used rejection sampling approaches adopt a boxcar kernel (0/1) to differentiate between behavioral ("1") and non-behavioral ("0") solutions, and use full-dimensional updating in pursuit of the posterior parameter distribution. This approach might work well for low-dimensional problems (e.g. $d \leq 10$) but is not particularly efficient in high-dimensional parameter spaces which require partial (subspace) sampling to rapidly locate posterior solutions.

In this chapter, we have introduced DREAM$_{(\text{ABC})}$ to permit diagnostic inference of complex system models. This approach uses Metropolis-within-Gibbs simulation with DREAM [*Vrugt et al.*, 2008a, 2009a] to delineate the space of behavioral (posterior) models. Three different case studies involving a simple one-dimensional toy problem, a 20-dimensional mixture of bivariate distributions, and a 14-dimensional hydrologic model calibration problem illustrate that DREAM$_{(\text{ABC})}$ is about 3 - 1,000 times more efficient than commonly used ABC sampling approaches. This gain in sampling efficiency increases with dimensionality of the parameter

space.

Chapter 5

Diagnosing Nonstationary Watershed Behavior from Process-Based Model Evaluation using Approximate Bayesian Computation

This Chapter has been submitted to Water Resources Research. Citation: Sadegh, M., J.A. Vrugt, C. Xu, and E. Volpi (2015), Diagnosing Nonstationary Watershed Behavior from Process-Based Model Evaluation using Approximate Bayesian Computation, *Water Resources Research*, Under Review.

5.1 Introduction

In the past decades much progress has been made in the measurement, modeling, and statistical/probabilistic analysis of surface hydrologic systems. A common assumption in many published studies is that measured/simulated hydrologic variables vary around some constant mean value with constant variance and serial correlation structure [*Clarke*, 2007]. Recent work, however, has called into question this conventional "stationarity" paradigm [*Westmacott and Burn*, 1997, *Karl and Knight*, 1998, *Strupczewski et al.*, 2001, *McCabe and Wolock*, 2002, *Groisman et al.*, 2004, *Fu et al.*, 2004, *Lins and Slack*, 2005, *Svensson et al.*, 2005, *Alexander et al.*, 2006, *Hodgkins and Dudley*, 2006, *Xu et al.*, 2006, *Leclerc and Ouarda*, 2007, *Milly et al.*, 2008, *Villarini et al.*, 2009, *Kundzewicz*, 2011, *Stedinger and Griffis*, 2011, *Vogel et al.*, 2011, *Waage and Kaatz*, 2011, *Ishak et al.*, 2013, *Salas and Obeysekera*, 2014]. For example, *Strupczewski et al.* [2001] showed evidence of nonstationarity in the annual maximum (AM) flows of 39 Polish rivers during a 70 years period of study (1921-1990). *Leclerc and Ouarda* [2007] conducted a regional flood frequency analysis of 29 different watersheds located in the southeastern part of Canada and northeastern part of the United States and showed that explicit consideration of trends in hydrologic regimes significantly improves flood quantile estimates at ungauged sites. *Villarini et al.* [2009] examined the annual peak discharge records from 50 different stations in the continental United States and concluded that almost 50% of the studied watersheds exhibit statistically significant changes in the mean and variance of annual flood peaks. A study by *Vogel et al.* [2011] reported flood magnification factors larger than two for several regions in the USA. A similar study by *Ishak et al.* [2013] used 30 to 97 years of AM streamflow data from Australian catchments and showed a significant downward trend in the AM flood values.

To help detect temporal trends, and nonstationary behavior of hydro-climate systems a host of different statistical/nonstatistical methods have been described and used in the water resources literature. Examples include the von Neuman ratio analysis [*Buishand*, 1982],

(flood) frequency and risk analysis [*Cunnane*, 1988, *Jain and Lall*, 2001, *Cunderlik and Burn*, 2003, *Cunderlik and Ouarda*, 2006, *Leclerc and Ouarda*, 2007, *Dettinger*, 2011, *Stedinger and Griffis*, 2011, *Chebana et al.*, 2013], spectral analysis [*Ramachandra Rao and Yu*, 1986, *Joshi and Pandey*, 2011, *Ni et al.*, 2011], moving average analysis [*Ramachandra Rao and Yu*, 1986, *Anghileri et al.*, 2014], Kendall, Mann-Kendall, Spearman, Pearson and Pettit tests [*Lins and Slack*, 1999, *Douglas et al.*, 2000, *Yue et al.*, 2002, *Zhang et al.*, 2009, *Bassiouni and Oki*, 2013, *Rougé et al.*, 2013, *Westra et al.*, 2013], wavelet analysis [*Karthikeyan and Nagesh Kumar*, 2013], generalized extreme value (GEV) models [*Clarke*, 2002, *El-Adlouni*, 2007, *Nasri et al.*, 2013, *Westra et al.*, 2013], and max-stable theory [*Westra and Sisson*, 2011].

The work of *Lins* [1985] is the first study on long-term analysis of hydrologic trends. This analysis showed that broad, national patterns in streamflow are fairly closely related to annual (continental US average) precipitation anomalies, especially during the post 1950 period for which the most extensive stream gage network exists. *Lettenmaier et al.* [1994] analyzed spatial patterns in trends of hydro-climatological variables and concluded that the observed trends in streamflow are not entirely consistent with the changes in temperature and precipitation and maybe due to a combination of climatic and water management effects. Many later studies have focused on the identification and interpretation of temporal changes in river flow regimes [*Zhang et al.*, 2001, *Burn and Hag Elnur*, 2002, *Birsan et al.*, 2005, *Novotny and Stefan*, 2007, *Petrow and Merz*, 2009, *Ishak et al.*, 2013], among many others. These studies employ a range of parametric and nonparametric approaches to detect temporal changes in hydrologic time series. *Khaliq et al.* [2009] presented a comprehensive review of the methodologies adopted for the identification of hydrologic trends, while *Kundzewicz and Robson* [2004] provided general guidance for the detection of trends in hydrologic data.

In a separate line of research, *Renard et al.* [2006] proposed a Bayesian framework to take into account the modeling uncertainties associated with nonstationary systems. Their study demonstrated significant improvements to the prediction of extreme events by using flexible

probability density functions whose parameters can be adapted to shifts (step-changes) and linear trends. Such time-dependant stochastic modeling is advocated in studies dating back as late as *North* [1980] and more recently, *Westerberg et al.* [2011], yet is difficult to be used outside the "calibration" data period. *Lima and Lall* [2010] developed hierarchial Bayesian models in a spacial scaling framework which can simulate trends of annual and monthly maximum streamflow, and reduce associated parameter and model uncertainties. *Ouarda and El-Adlouni* [2011] also employed a Bayesian framework and used prior information regarding the study variables for nonstationary frequency analysis. Finally, *Nasri et al.* [2013] developed a GEV model with B-Spline to model nonstationary meteorological processes. They conducted MCMC simulation in a Bayesian context to estimate the posterior distribution of the GEV-B-Spline model parameters and hydro-climate variable quantiles.

The term nonstationarity is a convenient way to describe characteristics of hydrologic time series and/or variables that do not meet the "constancy" criteria referred to above. Yet, this term fails to distinguish between long-term fluctuations (such as those caused by slowly-varying climate) and gradual/abrupt changes due to human intervention (such as regulation through reservoirs and dams, land-use changes, etc.). For instance, much of the downward trend in the AM flood peaks observed by *Ishak et al.* [2013] is not due to changes in watershed behavior (time invariant properties), but rather due to large-scale climate variability (different climate modes). Sudden multi-year changes in catchment behavior (single or multi-point) are often of anthropogenic origin, whereas decadal variations in streamflow response can be explained by climate trends and/or variability [*Leopold*, 1968, *Graf*, 1977, *Pitman*, 1978, *Sauer et al.*, 1983, *Potter*, 1991, *Smith et al.*, 2002, *Zhang and Schilling*, 2006, *Milly et al.*, 2008, *Vaze et al.*, 2010, *Villarini and Smith*, 2010].

Although trends can be readily apparent in time series of hydrologic variables, one should be particularly careful to immediately classify this behavior as nonstationarity. For instance, several published studies consider variations in hydro-climatic variables to be nonstationary

behavior [*Clarke*, 2002, *Parey et al.*, 2007, *Ishak et al.*, 2013, *Anghileri et al.*, 2014], others argue they might be explained by the natural variability of the system [*Kundzewicz et al.*, 2005, *Koutsoyiannis*, 2006, 2011, *Lins and Cohn*, 2011, *Koutsoyiannis*, 2013], also referred to as long term persistency [*Hurst*, 1951]. These long-term decadal fluctuations have been studied in detail by *Hurst* [1951] using discharge data from the Nile river in Egypt. This work attributed the short-term (up to about 100 years) trends in streamflow data to long term-persistency of the underlying watershed, rather than nonstationarity.

Without a sufficiently long data record it is particularly difficult to determine whether these temporal changes in watershed behavior are due to natural fluctuations in the weather and climate, or whether the watershed itself is experiencing changes, the latter which we classify as nonstationarity. As an illustrative example, consider a relatively simple model which superimposes twenty different stochastic processes with mean value drawn from a normal distribution with mean 3 and standard deviation 0.2. Each process lasts for a period of 100 years. The synthetic time series is then perturbed with white noise using a standard deviation of 0.1. Figure 5.1 plots the artificial data created by this model. The top panel (a) plots the first 200 years of the synthetic time series (blue line) with an average of roughly 3.15 (red line) for the first 100 years, and 2.70 (red line) for the remaining 100 years. Statistical and graphical analysis of this 200-year data set suggest the presence of nonstationarity. Indeed after 100 years, the system response abruptly shifts from one mode to the next. Although this sudden drift in system response/behavior might seem rather unrealistic for real-world situations, this example serves to illustrate the importance of data length. Indeed, if one views the behavior of this system over a much larger 2,000-year period, depicted in the bottom panel, then stationary behavior is observed. Thus, one should be particularly careful in proclaiming nonstationarity based on a relatively short data time period.

Many contributions to the hydrologic literature have used statistical analysis and/or visual interpretation of the observed system response to classify the watersheds behavior as station-

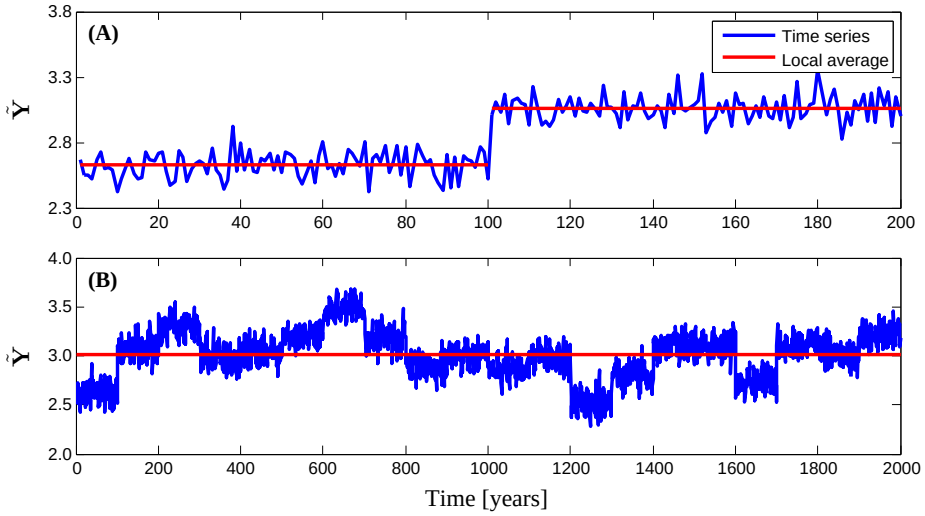


Figure 5.1: Synthetic time series of data generated by a stochastic stationary process to demonstrate the importance of considering time scales in nonstationary behavior analysis. The top panel provides a snapshot of 200 years of data from the original time series of 2,000 years (bottom panel). Artificial data is represented with a blue line, whereas the red line represents the mean system response. The top panel demonstrates a sudden shift in system response after exactly 100 years of observation. This abrupt change of the system response would seem a textbook example of nonstationary, yet when the behavior of the system is viewed over a much longer, 2,000 year data period, the mean and standard deviation of $\tilde{\mathbf{Y}}$ appear stable, and indeed stationarity is observed. One should therefore be particularly careful proclaiming nonstationarity on the basis of graphical and/or simple statistical analysis, and a relatively small data set. Note, this example is adapted from *Koutsoyiannis* [2011].

arity or not. Without further investigation into the underlying process physics, such approach can lead to spurious conclusions [*Koutsoyiannis*, 2006, 2011]. For instance, consider the top panel (a) of Figure 5.2 which plots (in red) the 10-year cumulative streamflow (2002-2012) of the Blackberry creek basin at Yorkville, IL. The dashed black and grey lines depict the slope of the cumulative streamflow data before and after 2007, respectively. If the response of this watershed to rainfall is deemed stationary, one would expect a somewhat constant slope in the period from 2002 to 2012. However, we observe a sharp transition in 2007 after which the rate of outflow (cumulative streamflow) has suddenly increased. This suggests an abrupt change in watershed behavior. This curve alone is insufficient to conclude the presence of nonstationarity (as suggested by `http://non-stationarities.irstea.fr/`). Indeed, the increase in streamflow volume can be explained by an intensifying hydrologic cycle with precipitation amounts that have significantly increased after 2007 (see bottom

panel). This finding reinforces the study of *Groisman et al.* [2001] which showed a "significant relationship" between the frequency of heavy precipitation and high streamflow for the eastern United States. Our diagnostic analysis presented in Section 5.3 of this chapter indeed confirms that the behavior of the Blackberry watershed is considered stationary during the 2002-2012 observational period. A similar conclusion was drawn by *Chong et al.* [2009] using data from the Yellow River in China. Streamflow trends were explained by large-scale decadal variations in climate forcing, and land use changes were found to have a negligible impact.

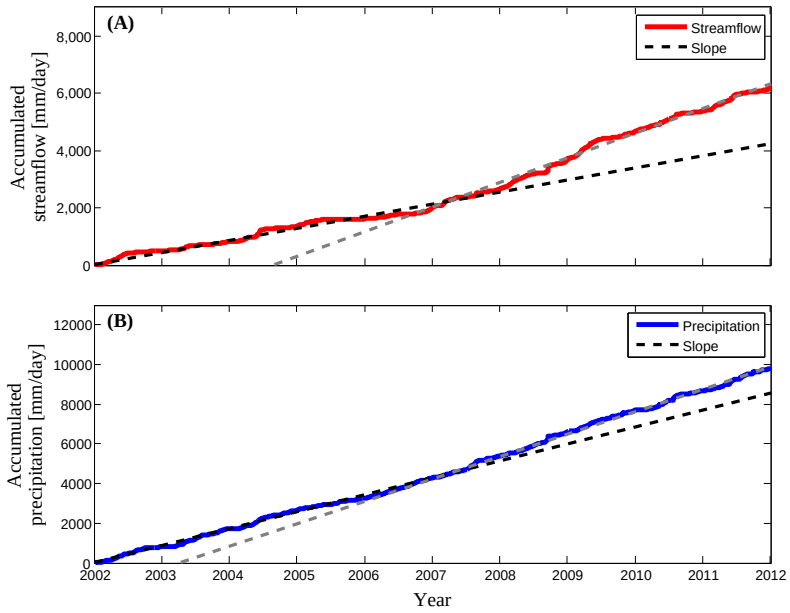


Figure 5.2: Cumulative streamflow (a) and precipitation (b) data record of the Blackberry creek in Illinois, USA for the years 2002-2012. The slope of the cumulative streamflow curve increases noticeably between the years 2007 and 2008. In the hydrologic community this behavior is classified as nonstarionary due to growing urbanization (`http://non-stationarities.irstea.fr/`). Yet, the bottom graph demonstrates noticeably larger precipitation amounts from 2007 onwards which coincides perfectly with the increase in streamflow. With the help of diagnostic model evaluation and temporal analysis of the hydrologic signatures, we classify this watershed as stationary. A single model structure and fixed parameterization is sufficient to model the rainfall-runoff response.

The premise of this chapter is that a process-oriented methodology with its roots in diagnostic model evaluation and Bayesian inference, can help to differentiate between stationary and nonstationary catchment behavior. This methodology follows recommendations of *Kout-*

soyiannis [2006] and *Koutsoyiannis* [2011], and uses diagnostic inference with approximate Bayesian computation (ABC) and summary metrics to pinpoint temporal variations in catchment behavior. These summary metrics can be viewed as fingerprints of the watershed, and their variation over time would convey subtle changes in catchment behavior. Indeed, if the watersheds behavior is stationary, then one would expect the summary metrics not to vary much over time, and a fixed model structure and single parameterization should be sufficient to adequately simulate the observed (time-invariant) rainfall-runoff transformation. Changes to the watersheds characteristics (properties) should affect signatures of the rainfall-runoff transformation, and induce temporal variations in the summary metrics. These variations cannot be simulated properly with a single model structure and fixed parameterization, and diagnostic model evaluation with ABC cannot find "proper" behavioral simulations.

The remainder of this chapter is organized as follows. In Section 5.2, we discuss the watersheds and data used herein. These watersheds have been carefully selected based on literature review and encompass stationary and nonstationary behavior. Then in Section 5.3, we shortly review the diagnostic model evaluation framework of *Vrugt and Sadegh* [2013], which is ideally suited to help detect hydrologic nonstationarity. This section proceeds with a discussion of the summary metrics used to quantify watershed behavior. In this section, we are particularly concerned with the choice of the cutoff threshold, ϵ to differentiate between behavioral and nonbehavioral models. Section 5.4 presents the results of our diagnostic analysis. Finally, Section 5.5 concludes this chapter with a summary of our most important findings.

5.2 Hydrologic Data and Watersheds

In this study, we analyze ten (assumed to be) stationary watersheds from the MOPEX data set [*Schaake et al.*, 2006], and nine (claimed to be) nonstationary watersheds. Table 5.1 lists

the name of each watershed, its geographical location (state, country), and catchment size (km^2). These watersheds range from 0.94 - 2,447 km^2 and are located in Australia and the USA.

The classification of each watershed as stationary or nonstationary appears in the literature and is based on visual interpretation of the discharge response, time-series analysis, and/or knowledge of human activity (urbanization, deforestation, etc.). For instance, the MOPEX watersheds used herein are assumed to be stationary, whereas the Wights catchment (deforestation after 3 years from commencement of monitoring in 1974, e.g. [*Mroczkowski et al.*, 1997]), Axe creek and Wimmera river basins (changing rainfall characteristics and much higher temperatures during 1997-2008), Flinders and Gilbert watersheds (cyclonic rainfall), and Ferson and Blackberry creek basins (urbanization) are assumed to be nonstationary. To verify whether our diagnostic analysis can correctly distinguish between stationary and nonstationary watersheds, we also include two artificial catchments with 10-year of synthetic daily streamflow observations created from the SAC-SMA model using forcing data from the French Broad river basin in the USA. Nonstationarity is enforced by either abruptly changing the SAC-SMA model parameterization from one day to the next in year 5 of the 10-year data period (watershed "abrupt"), or by linearly adapting over a 365-day period between year 5 and 6 the SAC-SMA parameterization (watershed "gradual"). Detailed information about the watersheds that are classified as nonstationary can be found at `http://non-stationarities.irstea.fr/`.

5.3 Approximate Bayesian Computation

ABC was introduced by *Diggle and Gratton* [1984] to permit inference of complex models in which the likelihood is intractable, computationally too expensive to evaluate or not explicitly available. The premise behind ABC is that model parameter $\boldsymbol{\theta}^*$ should be a sample from

the posterior distribution if distance between summary metrics of the model simulation and system response, $\rho\big(S(\mathbf{Y}(\boldsymbol{\theta}^*)), S(\tilde{\mathbf{Y}})\big)$, is less than some small positive threshold, ϵ [*Beaumont et al.*, 2002, *Marjoram et al.*, 2003, *Sisson et al.*, 2007, *Del Moral et al.*, 2008]. The degree of approximation of the posterior distribution, $p\big(\boldsymbol{\theta}|\rho\big(S(\mathbf{Y}), S(\tilde{\mathbf{Y}})\big) \leq \epsilon\big)$, depends on the choice of ϵ. The posterior parameter distribution will converge to the true distribution in the limit of $\epsilon \to 0$, pending that the summary metrics of the data are sufficient [*Pritchard et al.*, 1999, *Beaumont et al.*, 2002, *Turner and van Zandt*, 2012]. Sufficient summary metrics, $S(.)$, contain as much information as the original data itself. However, obtaining sufficient metrics is very difficult in practice, and loss of information in projecting the original data into one or more (lower dimensional) metrics inadvertently introduces an extra degree of approximation of the posterior distribution.

Table 5.1: Description of the watersheds used herein including name, region, USGS ID and drainage area (km^2), mean annual precipitation, $\tilde{\mathbf{P}}$ (mm/day), mean annual evapotranspiration, $\mathbf{P\tilde{E}T}$ (mm/day), and mean annual runoff coefficient, $\tilde{\mathbf{Y}}/\tilde{\mathbf{P}}$ (-). The top ten watersheds are derived from the MOPEX data set and deemed stationary in the hydrologic literature. The other nine watersheds are classified as nonstationary. Five of these watersheds are located in Australia, two are located in the United States, and the other two constitute synthetic data sets created with the SAC-SMA model and used to validate the proposed methodology. The last column lists, for each watershed, the minimum distance between the observed and simulated summary metrics (outcome of Equation 5.5).

Name	City (Region), State, Country	USGS ID	Area	$\tilde{\mathbf{P}}$	$\mathbf{P\tilde{E}T}$	$\tilde{\mathbf{Y}}/\tilde{\mathbf{P}}$	$\min(\rho(\cdot))$
		Stationary Watersheds					
Oostanaula river basin	Resaca, GA, USA	02387500	4,150	3.8281	2.3711	0.4218	0.0282
Pearl river basin	Edinburg, MS, USA	02482000	2,341	3.4510	2.8813	0.3010	0.0426
Bogue Chitto river basin	Bush, LA, USA	02492000	3,141	3.4510	2.8813	0.3010	0.0350
Little Pigeon river basin	Sevierville, TN, USA	03470000	914	3.8589	2.9268	0.3384	0.0361
Tangipahoa river basin	Robert, LA, USA	07375500	1,673	3.6837	2.1868	0.4065	0.0222
Comite river basin	Comite, LA, USA	07378000	735	3.9523	2.9356	0.3767	0.0270
Calcasieu river basin	Oberlin, LA, USA	08013500	1,950	3.7730	2.9496	0.3538	0.0449
S. Umpqua river basin	Tiller, OR, USA	14308000	1,162	4.0382	2.9990	0.3559	0.0222
French Broad river basin	Asheville, NC, USA	03451500	2,447	4.2034	2.2440	0.4858	0.0155
Leaf river basin	Collins, MS, USA	02472000	1,924	3.5496	2.9562	0.3105	0.0216
		Nonstationary Watersheds					
Axe creek basin	Longlea, Victoria, Australia	N/A	1,235	1.5589	3.3371	0.0329	0.1706
Wimmera river basin	Glenorchy Concrete Weir Tail, Victoria, Australia	N/A	2,000	1.5620	3.1558	0.0663	0.0667
Wights basin	Collie, Western Australia	N/A	0.94	3.0232	3.8398	0.2309	0.1018
Flinders river basin	Glendower, Queensland, Australia	N/A	1,911	1.9541	4.6474	0.0914	0.2525
Gilbert river basin	Gilberton, South Australia	N/A	1,906	2.3679	4.6475	0.1263	0.1676
Ferson creek basin	St. Charles, IL, USA	05551200	134	2.7380	2.0553	0.2909	0.0306
Blackberry creek basin	Yorkville, IL, USA	05551700	181	2.6809	2.1013	0.2855	0.0309
Synthetic Case I ("abrupt")	—	—	—	4.2034	2.2440	0.5177	0.1220
Synthetic Case II ("gradual")	—	—	—	4.2034	2.2440	0.5165	0.1443

The use of summary metrics holds three great promises for diagnostic model evaluation. In the first place, the use of summary metrics rather than some convoluted series of error residuals, helps ensure that the model accurately mimics properties of system behavior observed in the data and deemed important to the user. Secondly, and perhaps most importantly, if we carefully design each summary metric to be sensitive to one component (process) of the model, then it is easy to pinpoint which part of the model is malfunctioning. Finally, summary metrics can be designed so that they are relatively insensitive to errors in the forcing data (precipitation data errors). This helps with decomposition of the error residuals into its constituent sources. If the summary metrics are sufficient and thus contain as much information as the original calibration data itself, then the ABC-derived posterior distribution is very similar, if not identical, to the "true" posterior distribution.

Figure 5.3 schematically depicts the ABC framework. First, the user defines one or more summary metrics, $S(\tilde{\mathbf{Y}})$ that summarize the original data in a lower dimensional space, and specifies a value for ϵ which is used to discriminate among behavioral (posterior) and non-behavioral samples. Then, N samples, $\{\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_N^*\}$ are drawn from the prior parameter distribution. In the absence of detailed prior information about the parameters, and/or lack of useful regionalization relationships, this distribution is often assumed to be uniform in environmental modeling studies (note *Scharnagl et al.* [2011] present a strategy to assess informative prior distributions in vadose zone modeling). Each proposal is subsequently evaluated by the model, and used to calculate the simulated summary metrics, $S(\mathbf{Y}(\boldsymbol{\theta}_i^*))$, where $i = 1, \ldots, N$. Simulated metrics are then compared to their observed counterparts using some distance function, $\rho(\cdot)$. If $\rho(S(\mathbf{Y}(\boldsymbol{\theta}_i^*)), S(\tilde{\mathbf{Y}})) \leq \epsilon$ then $\boldsymbol{\theta}_i^*$ is considered to be a posterior sample. Finally, the accepted samples are used to summarize the target distribution, $p(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$.

Obviously, the degree of approximation of the true posterior distribution depends in large part on the chosen summary statistics, in addition to the value of ϵ. As sufficient metrics

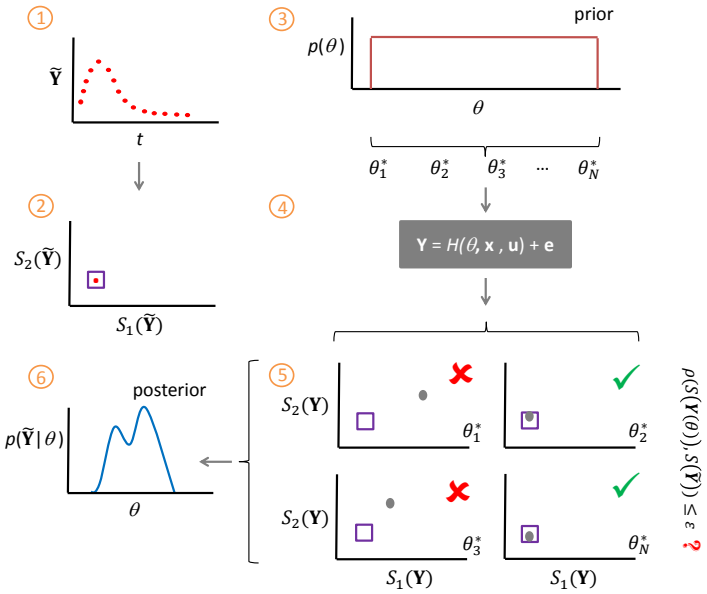


Figure 5.3: Schematic overview of likelihood-free (ABC) inference for a model involving only one parameter using two different summary metrics, S_1 and S_2 of the measured data, $\tilde{\mathbf{Y}}$. The original watershed data, $\tilde{\mathbf{Y}}$ is plotted in step 1, and is used to calculate the observed summary metrics (step 2). After the user has defined the value of ϵ, step 3 proceeds by drawing N samples from the prior parameter distribution, $p(\theta)$. In step 4, each of these N samples is evaluated by the model, and used to calculate the simulated values of the summary metrics. Step 5 then proceeds with a comparison of the observed and simulated summary metrics. Model realizations whose simulated summary metrics fall within the purple box, and thus satisfy $\rho(S(\mathbf{Y}(\boldsymbol{\theta}_i^*)), S(\tilde{\mathbf{Y}})) \leq \epsilon$ are used in step 6 to approximate the posterior parameter distribution.

are practical exceptions for complex systems, the choice of the summary metrics has to be done with care - also keeping in mind the purpose of application. Section 5.3.2 will elucidate our choice of the summary metrics in the present application of diagnostic model evaluation to the analysis of watershed nonstationary in response to rainfall.

5.3.1 Posterior sampling: DREAM$_{(\mathrm{ABC})}$

Application of likelihood-free inference with ABC requires the availability of a sampling method that can efficiently search the parameter space in pursuit of behavioral solutions. Commonly used rejection sampling methods are rather inefficient, and we therefore take advantage of the DREAM$_{(\mathrm{ABC})}$ algorithm developed by *Sadegh and Vrugt* [2014]. This general-purpose methodology combines the strengths of Markov chain Monte Carlo (MCMC)

simulation, differential evolution, and subspace sampling to rapidly explore the behavioral parameter space. Previous work has shown that DREAM$_{(\text{ABC})}$ is several orders of magnitude more efficient than rejection based sampling methods such as ABC-REJ and ABC-PMC [*Sadegh and Vrugt*, 2014].

In DREAM, K ($K > 2$) different Markov chains are run simultaneously in parallel, and multivariate proposals are generated on the fly from the collection of chains, $\boldsymbol{\Theta}^{t-1}$ (matrix of $K \times d$ with each chain state as row vector) using differential evolution [*Storn and Price*, 1997, *Price et al.*, 2005]. If A is a subset of δ-dimensional space of the original parameter space, $\mathbb{R}^{\delta} \subseteq \mathbb{R}^{d}$ then a jump in the kth chain, $k = \{1, \ldots, K\}$ at iteration $t = \{2, \ldots, T\}$ is calculated using

$$\begin{aligned} \boldsymbol{\Delta}^{*}_{k,A} &= (\mathbf{1}_{\delta} + \boldsymbol{\lambda}_{\delta})\gamma_{(\tau,\delta)} \left[\sum_{j=1}^{\tau} \boldsymbol{\theta}^{t-1}_{\mathbf{r}^1_j,A} - \sum_{j=1}^{\tau} \boldsymbol{\theta}^{t-1}_{\mathbf{r}^2_j,A} \right] + \boldsymbol{\zeta}_{\delta} \\ \boldsymbol{\Delta}^{*}_{k,\neq A} &= 0, \end{aligned} \tag{5.1}$$

where $\gamma = 2.38/\sqrt{2\tau\delta}$ is the jump rate, τ denotes the number of chain pairs used to generate the jump, and $\mathbf{r}^1$ and $\mathbf{r}^2$ are τ-vectors with integer values drawn without replacement from $\{1, \ldots, k-1, k+1, \ldots, K\}$. The values of $\boldsymbol{\lambda}_{\delta}$ and $\boldsymbol{\zeta}_{\delta}$ are sampled independently from the uniform distribution $\mathcal{U}_{\delta}(-c, c)$ and normal distribution $\mathcal{N}_{\delta}(0, c^{*})$ with, typically, $c = 0.1$ and c^{*} small compared to the width of the target distribution, $c^{*} = 10^{-12}$ say.

The candidate point of chain k at iteration t then becomes

$$\boldsymbol{\theta}^{*}_{k} = \boldsymbol{\theta}^{t-1}_{k} + \boldsymbol{\Delta}^{*}_{k}, \tag{5.2}$$

and a modified selection rule is used to determine whether to accept this proposal or not.

This selection rule is defined as,

$$P_{\text{acc}}(\boldsymbol{\theta}_k^{t-1} \rightarrow \boldsymbol{\theta}_k^*) = \begin{cases} I\big(f(\boldsymbol{\theta}_k^*) \geq f(\boldsymbol{\theta}_k^{t-1})\big) & \text{if } f(\boldsymbol{\theta}_k^*) < 0 \\ 1 & \text{if } f(\boldsymbol{\theta}_k^*) \geq 0 \end{cases}, \tag{5.3}$$

where the function, $f(\cdot)$ is calculated as follows

$$f(\boldsymbol{\theta}) = \epsilon - \rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}))\big), \tag{5.4}$$

and the distance function is defined as

$$\rho\big(S(\mathbf{Y}(\boldsymbol{\theta}), S(\tilde{\mathbf{Y}}))\big) = \max_{i=1:l} \left(\big| S_i(\tilde{\mathbf{Y}}) - S_i(\mathbf{Y}(\boldsymbol{\theta})) \big| \right). \tag{5.5}$$

Thus, we accept, $P_{\text{acc}}(\boldsymbol{\theta}_k^{t-1} \rightarrow \boldsymbol{\theta}_k^*) = 1$, if the fitness of $\boldsymbol{\theta}_k^*$ is larger than that of the current state of the kth chain, $\boldsymbol{\theta}_k^{t-1}$ or if the summary metrics of the proposal are within ϵ of their observed counterparts, $f(\boldsymbol{\theta}_k^*) \geq 0$, otherwise the candidate point is rejected. After a burn-in period in which $f(\cdot) \geq 0$, the convergence of DREAM$_{\text{(ABC)}}$ can be monitored with the $\hat{R}$ diagnostic of *Gelman and Rubin* [1992].

For all watersheds we use the default settings of DREAM$_{\text{(ZS)}}$, and set $K = 3$ in combination with a maximum of $100,000$ model evaluations. This equates to $T = 33,333$ samples in each of the three chains.

5.3.2 Choice of Summary Metrics

Likelihood-free inference requires the user to specify one or more summary statistics which properly extract the information content of the available data. The choice of which metrics to use depends on the goal of ABC application. If the main subject of interest is posterior

approximation, then the summary statistics should be sufficient and "convey all relevant information" (quote of *Edwards* [1992]) about the data. In other words, the chosen metrics should contain as much information as the original data itself. Unfortunately, complex system models rarely admit sufficient statistics, let alone, the probabilistic properties of $\mathbf{e}$ should closely mimic their true behavior.

Self-sufficiency is not necessarily a requirement for diagnostic model evaluation in which the goal is not to approximate the "true" posterior distribution but rather to detect epistemic errors arising from incomplete and/or inadequate process knowledge. The advantage of summary statistics is that they exhibit a much better diagnostic power than some convoluted set of error residuals, and if properly designed and rooted in the relevant environmental theory, can help to illuminate to what degree a representation of the real world has been adequately achieved and how the model (null hypothesis) should be refined.

We now have to decide which summary statistics to use. To facilitate diagnostic inference of watershed models, it would be highly desirable if the chosen summary statistics are, (a) properly rooted in hydrologic theory, (b) independent (uncorrelated), (c) invariant to precipitation data errors, and (d) sensitive only to one specific component (process/equation) of the model. For the scope of the present chapter, characteristics (a), (b) and (c) are required to avoid proclaiming nonstationarity for the wrong reasons. We will elaborate on this in the next paragraph. If the goal of diagnostic analysis is to help detect epistemic errors then characteristic (d) is of crucial importance. Indeed, a mismatch between the measured and simulated summary metrics is then easily linked to an individual process in the model. This simplifies the analysis of model malfunctioning.

If the summary metrics are designed so that they extract the relevant signatures of catchment behavior, then analysis of their temporal dynamics (values for different time periods) can help elucidate subtle changes in the basin response to rainfall. In case that the summary metrics vary around some constant mean, a model structure with a single parameterization

suffices to adequately mimic the observed catchment behavior. However, if the summary statistics exhibit a sudden change (or drift) then this behavior is impossible to adequately simulate with a fixed model parameterization. The presence of such temporal trends is symptomatic for catchments that have experienced a sudden/steady rainfall characteristic change or undergone alterations due to for instance urbanization, deforestation, and/or changes in land-use. For this analysis to work well in practice, it is highly desirable that the summary statistics are independent. Indeed, nonstationarity might not be visible in all signatures of catchment response, and hence it is important to use multiple different summary metrics that each measure different and complementary parts of system behavior. Finally, to rule out the effect of forcing data errors, we advise to use summary metrics that are insensitive to precipitation data errors (timing + amounts). We certainly would not want to proclaim watershed nonstationarity if rainfall data errors are the culprit.

Unfortunately it is not particularly easy to formulate summary metrics that satisfy each of the three (possibly four) requirements listed above. We therefore resort to commonly used hydrologic criteria proposed in the literature. Table 5.2 lists each summary metric used herein along with a reference of how they are calculated from the observed rainfall and/or discharge record. In short, our analysis includes the following criteria: the baseflow index and runoff coefficient, fitting parameters of the flow duration curve (FDC) [*Vrugt and Sadegh*, 2013], slope of the FDC [*Yadav et al.*, 2007], rising and declining limb densities of the hydrograph [*Shamir et al.*, 2005], mean, median, coefficient of variation, skewness and maximum range of the observed daily flows, and maximum and minimum monthly and yearly flows, low and high pulse count and duration, flood duration and frequency, rise and fall rate, number of zero flow days [*Olden and Poff*, 2003], and peak distribution [*Euser et al.*, 2013].

Table 5.2: Name and description of each of the summary statistics used in our rainfall error analysis. The last column provides a reference for each metric, and can be used as guide for their numerical calculation.

Name	Description	Reference
Baseflow Index	Ratio between the total baseflow and streamflow volumes in the period of study	*Eckhardt* [2005]
Runoff Coefficient	Ratio between the total streamflow and precipitation volumes in the period of study	*Savenije* [1996]
FDC	Relationship between the exceedence probability of streamflow and its magnitude	*Searcy* [1959]
Slope of the FDC	Slope of the midpart of the FDC (between 33% and 66% exceedance probability rates)	*Yadav et al.* [2007]
Rising Limb Density (RLD)	Ratio between the number of peaks and cumulative time of rising limbs	*Morin et al.* [2002]
Declining Limb Density (DLD)	Ratio between the number of peaks and cumulative time of declining limbs	*Shamir et al.* [2005]
Mean Daily Flow (DF)	Average of daily streamflow values	*Clausen and Biggs* [2000]
Median DF	Median of daily streamflow values	*Clausen and Biggs* [2000]
Coefficient of Variation of DFs	Ratio between the standard deviation and mean of daily streamflow vlues	*Clausen and Biggs* [2000]
Skewness of DFs	Skewness of daily streamflow values	*Clausen and Biggs* [2000]
Ranges in DFs	Ratio of $10th/90th$, $20th/80th$, and $25th/75th$, percentiles of daily streamflow values	*Olden and Poff* [2003]
Mean Maximum Monthly Flow	Average of the maximum monthly streamflow values	*Olden and Poff* [2003]
Mean Minimum Monthly Flow	Average of the minimum monthly streamflow values	*Olden and Poff* [2003]
Median Maximum Monthly Flow	Median of the maximum monthly streamflow values	*Olden and Poff* [2003]
Median Minimum Monthly Flow	Median of the minimum monthly streamflow values	*Olden and Poff* [2003]
Low Flow Pulse Count	Number of events with streamflow values below $25th$ percentile	*Olden and Poff* [2003]
High Flow Pulse Count	Number of events with streamflow values above $75th$ percentile	*Olden and Poff* [2003]
Low Flow Pulse Duration	Mean duration of low streamflows	*Olden and Poff* [2003]
High Flow Pulse Duration	Mean duration of high streamflows	*Olden and Poff* [2003]
Flood Duration	Mean number of days that flow magnitude remains above flood threshold	*Olden and Poff* [2003]
Flood Frequency	Mean number of high flow (flood) events per year	*Olden and Poff* [2003]
Rise Rate	Average rate of positive flow changes from one day to the next	*Olden and Poff* [2003]
Fall Rate	Average rate of negative flow changes from one day to the next	*Olden and Poff* [2003]
Number of Zero Flow Days	Number of days with zero streamflow	*Olden and Poff* [2003]
Peak Distribution	Average slope between $10th$ and $50th$ percentiles of the FDC from peak flows only	*Euser et al.* [2013]

To discern which of these metrics is least sensitive to precipitation data errors we devised a simple numerical experiment with the SAC-SMA model [*Burnash et al.*, 1973]. Synthetic daily streamflow observations were created with ten different parameter combinations, drawn randomly from their uniform prior distribution (see Table 5.3), using forcing data from the French Broad river basin at Asheville, NC. For each of the ten streamflow simulations the summary metrics of Table 5.2 were calculated and considered to be our observed values. Then, a total of thousand different precipitation time series were created by perturbing the original rainfall record of the French Broad river basin with a 20% (heteroscedastic) measurement error. To also introduce errors in rainfall timing, individual storm events were randomly shifted with one day. The probability of such forward or backward shift was set to 20%. This leaves us with an ensemble of one thousand different hyetographs. The SAC-SMA was subsequently evaluated for each of the ten parameter combinations using the thousand different corrupted rainfall records. For each of those discharge simulations the summary metrics of Table 5.2 were calculated. To investigate the effect of data length on the simulated summary statistics, we consider discharge simulations of length 1, 2, 5, and 10-years.

Table 5.3: Description of the SAC-SMA model parameters and their (uniform) prior uncertainty ranges.

Parameter	Symbol	Minimum	Maximum	Units
Upper zone tension water maximum storage	UZTWM	1.0	150.0	mm
Upper zone free water maximum storage	UZFWM	1.0	150.0	mm
Lower zone tension water maximum storage	LZTWM	1.0	500.0	mm
Lower zone free water primary maximum storage	LZFPM	1.0	1000.0	mm
Lower zone free water supplemental maximum storage	LZFSM	1.0	1000.0	mm
Additional impervious area	ADIMP	0.0	0.40	-
Upper zone free water lateral depletion rate	UZK	0.1	0.5	day^{-1}
Lower zone primary free water depletion rate	LZPK	0.0001	0.025	day^{-1}
Lower zone supplemental free water depletion rate	LZSK	0.01	0.25	day^{-1}
Maximum percolation rate	ZPERC	1.0	250.0	-
Exponent of the percolation equation	REXP	1.0	5.0	-
Impervious fraction of the watershed area	PCTIM	0.0	0.1	-
Fraction percolating from upper to lower zone free water storage	PFREE	0.0	0.6	-

Figure 5.4 displays the results of our analysis, and plots histograms of the distance between the observed and simulated values of the summary statistics. The top panel (a-e) depicts the five metrics that are least sensitive to rainfall data errors, whereas the bottom panel (f-j) summarizes the results of the five most sensitive criteria. The results pertain to a 10-year data period and summarize the outcome of ten different SAC-SMA parameterizations. The distributions of the baseflow index, runoff coefficient, two fitting coefficients of the FDC, and annual rising limb density (RLD) are very well defined and exhibit a negligible dispersion. From all the different criteria listed in Table 5.2, these five summary metrics are least sensitive to precipitation data measurement errors. As the RLD is defined for each year separately, this equates to 14 different metrics for a 10-year data period. The other metrics of Table 5.2 exhibit considerable uncertainty and hence are not particularly useful for analysis of watershed nonstationarity. Similar conclusions are drawn for shorter simulation periods (not shown herein).

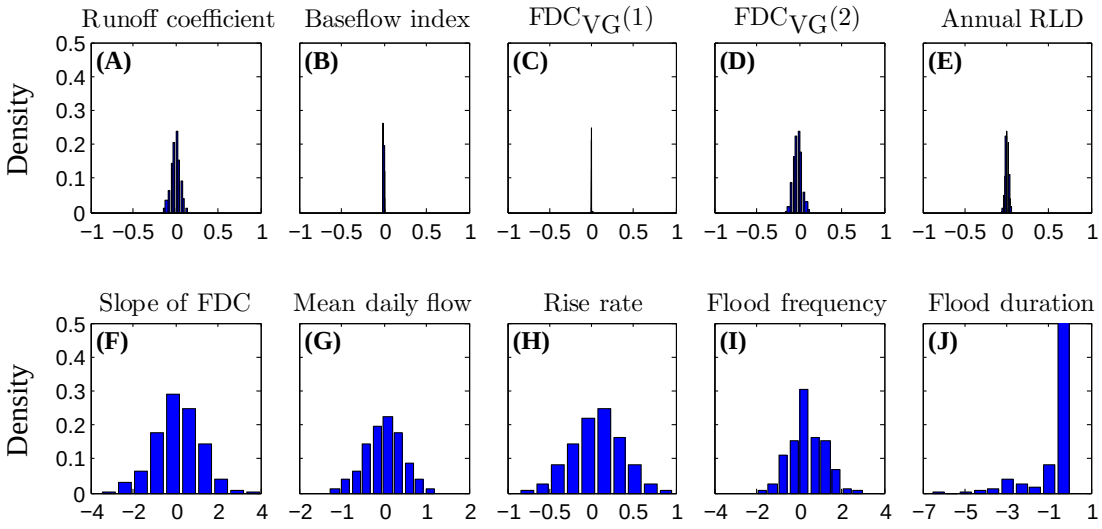


Figure 5.4: Histograms of the residuals between the observed and simulated summary metrics. The observed summary metrics are derived from ten different SAC-SMA discharge simulations each with a different parameterization. The simulated metrics are derived for the same SAC-SMA parameterizations but using thousand error corrupted rainfall records. The upper panel (a-e) plots the results for the least sensitive summary metrics, whereas the bottom panel (f-j) considers metrics that are much more sensitive to rainfall data errors.

5.3.3 Choice of ϵ, the Cutoff Threshold

Now our summary statistics have been defined we are left with selection of the value of the cutoff threshold ϵ used to differentiate between behavioral and non-behavioral models (simulations). If the goal of ABC application is to approximate the true posterior distribution, then the value of ϵ should be taken really small (e.g. 0.025). For diagnostic model evaluation, however, it is not immediately apparent what value of ϵ is deemed most appropriate for detection of epistemic errors. One approach is to derive the value of ϵ from analysis of the temporal variability of each summary metric. This approach ensures that the behavioral model space accurately portrays the observed properties of the data. For instance, one could assume ϵ to be equivalent to the standard deviation of each metric for different time periods. For watersheds that have undergone changes one would expect the value of ϵ derived this way to be larger than for watersheds that are deemed to exhibit stationary behavior. This is the underlying premise of the present chapter. Of course, care should be exercised that a sufficiently long data record is used to calculate the temporal dynamics of each summary statistics.

Table 5.1 (last column) lists the minimum values of $\rho(\cdot)$ for each of the 19 different watersheds derived from nonlinear minimization of Equation 5.5 using the SAC-SMA model and DREAM algorithm. Equation 5.5 is composed of the 14 different summary metrics selected in section 5.3.2. The tabulated results confirm our hypothesis. Indeed, the smallest values of $\rho(\cdot)$ are observed for watersheds that have been classified as stationary. Substantially larger values for $\rho(\cdot)$ are observed for the nonstationary watersheds, with the exception of the Ferson and Blackberry creek basins in the USA. Quite interestingly, our diagnostic analysis with ABC confirms our earlier findings in Figure 5.2 that these two basins are erroneously classified as nonstationary despite their significant urbanization. A value of $\epsilon = 0.05$ appears to be a good threshold to distinguish between stationary and nonstationary watersheds. In practice this means that if for a given watershed the SAC-SMA model cannot find distance

values smaller than 0.05, the basins properties or rainfall characteristics likely have experienced changes during the calibration data period. The resulting behavior is then classified as nonstationary. Thus, we now have a relatively simple process-based methodology which is rooted in diagnostic model evaluation and can hep to determine whether a basin exhibits stationary behavior or not.

It is important to note that the benchmark value of $\epsilon = 0.05$ to help differentiate between stationary and nonstationary watersheds is not a generic finding that applies to all rainfall-runoff simulators but rather to the SAC-SMA model or other models with similar structural complexity. If a more parsimonious model structure were used with fewer calibration parameters, then one would expect the value of ϵ to increase somewhat to compensate for larger errors in modeling the observed signatures. In the next section we illustrate the results of diagnostic model evaluation with ABC using data from three watersheds that have been classified as "stationary", and three basins (one real-world and two synthetic) that are known to exhibit "nonstationary" behavior.

5.4 Results and discussion

We now illustrate the results of our numerical simulations with the SAC-SMA model using the DREAM$_{(\mathrm{ABC})}$ algorithm with 14 different summary statistics (see section 5.3.2) derived from ten years of daily hydrologic data from the 19 different watersheds listed in Table 5.1. In our discussion, we focus attention on three stationary watersheds (Oostanaula, French Broad, and Leaf river basins in the United States) and three nonstationary watersheds (Wights catchment in Australia, and two synthetic data sets). The findings for these basins are representative for the entire collection of watersheds studied herein, and demonstrate the ability of the proposed methodology to detect subtle changes in the response of watersheds to rainfall, and hence to differentiate between stationary and nonstationary behavior.

Figure 5.5 presents histograms of the marginal posterior distribution of a representative set of five SAC-SMA parameters including (a) PCTIM, (b) ADIMP, (c) LZTWM, (d) LZPK, and (e) LZSK using the observed values of the 14 different summary metrics of the Oostanaula (top panel), French Broad (middle panel) and Leaf (bottom panel) river basins and $\epsilon = 0.05$. These three basins are classified as stationary in the hydrologic literature. The x-axis of each plot matches exactly the ranges of each parameter used in the (uniform) prior distribution (Table 5.3). The results in this Figure highlight several important findings. First, most of the SAC-SMA parameters are not particularly well resolved by calibration against the observed summary metrics. The marginal distributions encompass a large part of the prior distribution. The exception is LZPK which appears reasonably well constrained for each of the stationary watersheds. Second, most of the histograms are rather erratic and poorly described with a normal, log-normal, uniform or other traditional probability distributions. These results suggest that the posterior surface is non-smooth and rather difficult to sample. Similar conclusions were drawn in our earlier work presented in *Vrugt and Sadegh* [2013]. Finally, although the marginal distributions exhibit considerable scatter - multivariate plots of the behavioral solutions demonstrate that the posterior samples occupy only a small portion of the uniform prior hypercube (not shown).

To provide better insights into the sampled distance values of Equation 5.5, Figure 5.6 plots trace plots of the value of $\rho(\cdot)$ in each of the $K = 3$ Markov chains simulated with DREAM$_{(\text{ABC})}$. We illustrate the results for the (a) Oostanaula, (b) French Broad, and (c) Leaf river watersheds. These three watersheds are classified as stationary. Each of the three Markov chains is coded with a different color. The dashed red line represents the threshold of $\epsilon = 0.05$ which is used to differentiate between stationary and nonstationary watersheds. The DREAM$_{(\text{ABC})}$ algorithm rapidly converges to a limiting distribution. About 5,000 function evaluations are required for each watershed to locate behavioral solutions. The number of function evaluations with DREAM$_{(\text{ABC})}$ is more than sufficient to generate a large sample of the posterior distribution. Note that the different chains mix relatively well - indeed the

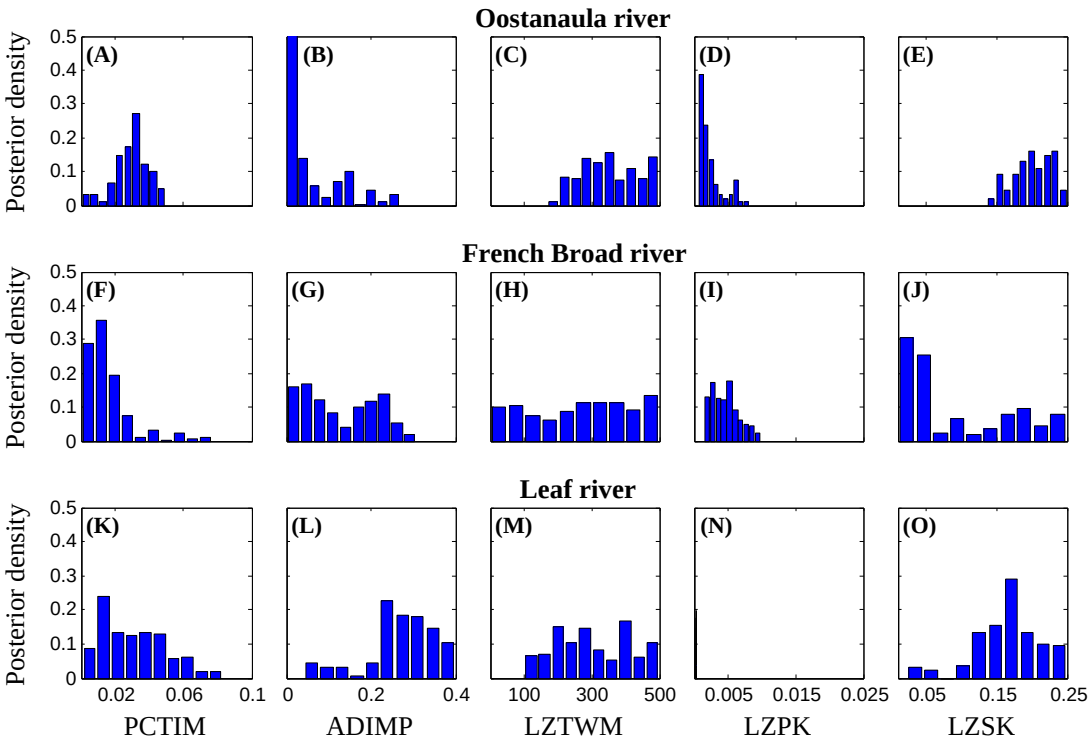


Figure 5.5: Marginal posterior parameter distributions of the SAC-SMA model parameters PCTIM, ADIMP, LZTWM, LZPK, and LZSK derived from DREAM$_{(ABC)}$ using historical data from the Oostanaula (top), French Broad (middle) and Leaf river (bottom) watersheds, respectively. Some parameters are very well defined by calibration against the 14 summary metrics, whereas others exhibit considerable uncertainty.

acceptance rate varies between 3.24 - 5.54%. Our diagnostic analysis confirms that the three watersheds indeed exhibit stationary behavior during the 10-year data period.

I now move on to the sampled summary statistics. Figure 5.7 plots histograms of the marginal distribution of the sampled summary statistics of the SAC-SMA model for the Oostanaula (top panel), French Broad (middle panel) and Leaf river (bottom panel) watersheds. The observed values of each of the summary statistics is indicated separately in each panel with a red cross. As it is particularly difficult to summarize the results of all the 14 summary statistics, we plot only the distributions of S_1, the runoff coefficient (a,f,k), S_2, the baseflow index (b,g,l), S_3 and S_4, the two fitting coefficients of the FDC function (c,h,m) and (d,i,n), and S_{11}, the RLD of year 7 of the simulation period (e,j,o). The sampled distributions of the summary metrics center nicely around their observed and appear rather uniform. As

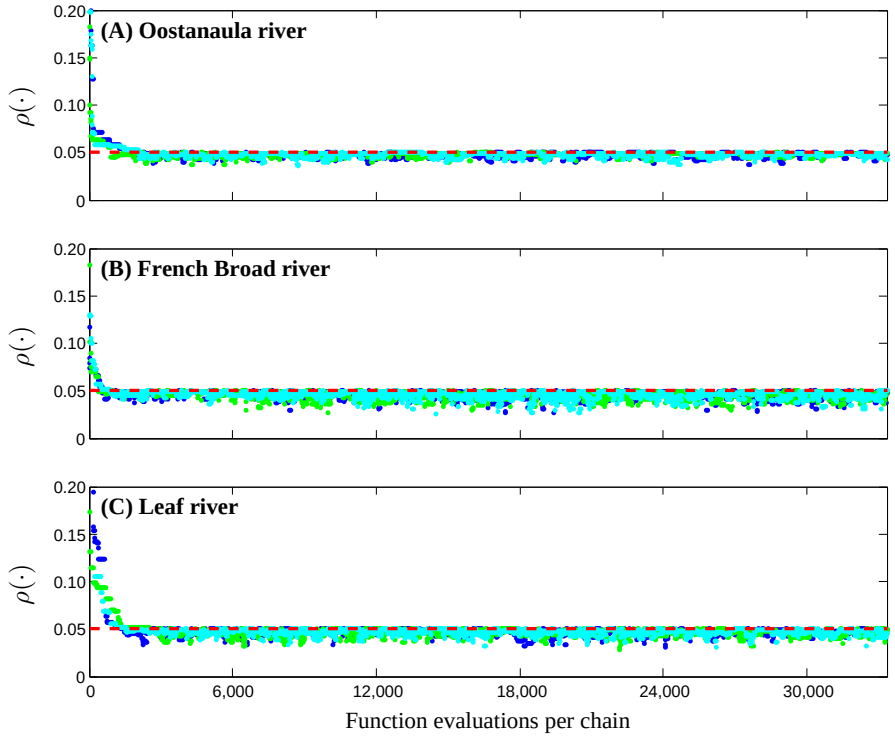


Figure 5.6: Trace plots of the sampled distance values, $\rho(\cdot)$, in each of the chains for the Oostanaula (a), French Broad (b) and Leaf (c) river watersheds. The dashed red line depicts the threshold of $\epsilon = 0.05$ used to differentiate between behavioral and nonbehavioral samples.

requested, all the sampled summary metrics reside within a distance of $\epsilon = 0.05$ of their observed values. Altogether, these results again illustrate the ability of the SAC-SMA model to portray accurately the rainfall-runoff response of the three different stationary watersheds. A fixed parameterization appears sufficient to mimic adequately close the observed summary statistics.

Figure 5.8 illustrates how the SAC-SMA posterior parameter uncertainty translates into streamflow simulation uncertainty. A representative 365 day period of the 10-year data record is used to illustrate our findings. The top panel displays the results for the Oostanaula river whereas the middle and bottom panels depict the results for the French Broad and Leaf river basins, respectively. The observed discharge data are indicated with red dots. The 95% streamflow uncertainty ranges appear rather narrow but nicely track the observed discharge data. About 40-60% of the discharge observations are contained in the 95% simulation

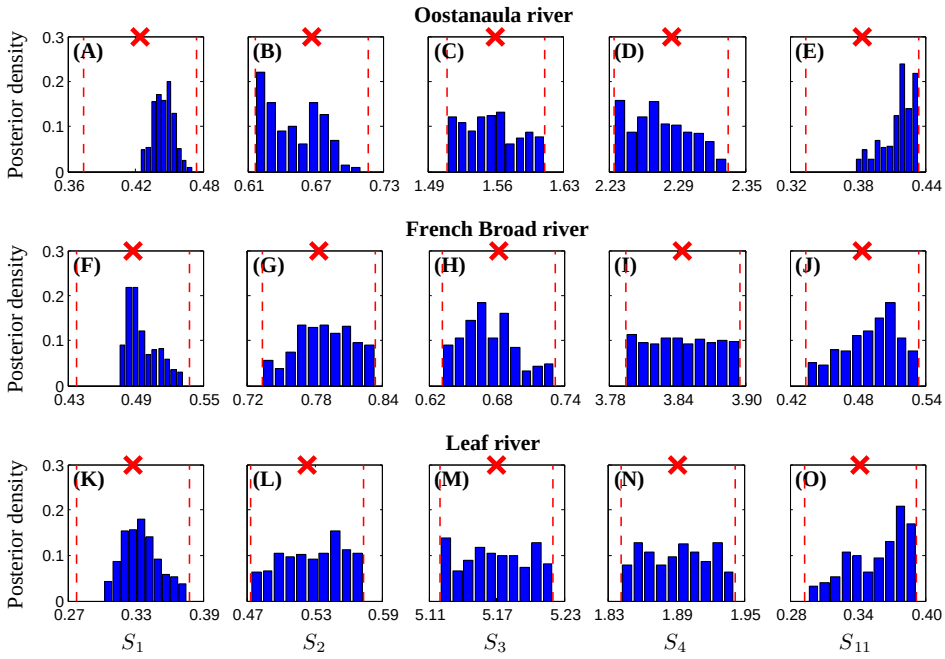


Figure 5.7: Histograms of the marginal posterior distributions of the runoff coefficient (S_1), baseflow index (S_2), FDC (S_3 and S_4) and rising limb density for the 7th year of the calibration period for the Oostanaula (up), French Broad (middle) and Leaf (bottom) river watersheds. The symbol "x" denotes the measured values of the summary metrics. These plots confirm that each of the posterior summary metrics lie within $\epsilon = 0.05$ (vertical dashed red lines) of their observed counterparts.

intervals. The Root Mean Square Error (RMSE) of the mean ensemble simulation ranges between 0.8 - 1.2 mm/day. This value would be substantially lower if a residual based likelihood function were used. This finding is not surprising (see also *Vrugt and Sadegh* [2013]) as our diagnostics approach does not fit directly against the observed discharge data, but rather is aimed at correctly representing the observed fingerprints of watershed behavior. A residual based fitting approach often leads to corrupted parameter values to counter measurement errors of the forcing data. This leads to a calibrated model that might best fit the observed discharge data but at the expense of a correct reproduction of the observed summary statistics of watershed behavior.

We now illustrate our findings for three of the watersheds that are categorized as nonstationary. Figure 5.9 presents trace plots of the sampled distance values, $\rho(\cdot)$ of Equation 5.5 in each of the $K = 3$ different Markov chains simulated with DREAM$_{(\text{ABC})}$ for the (a) Wights

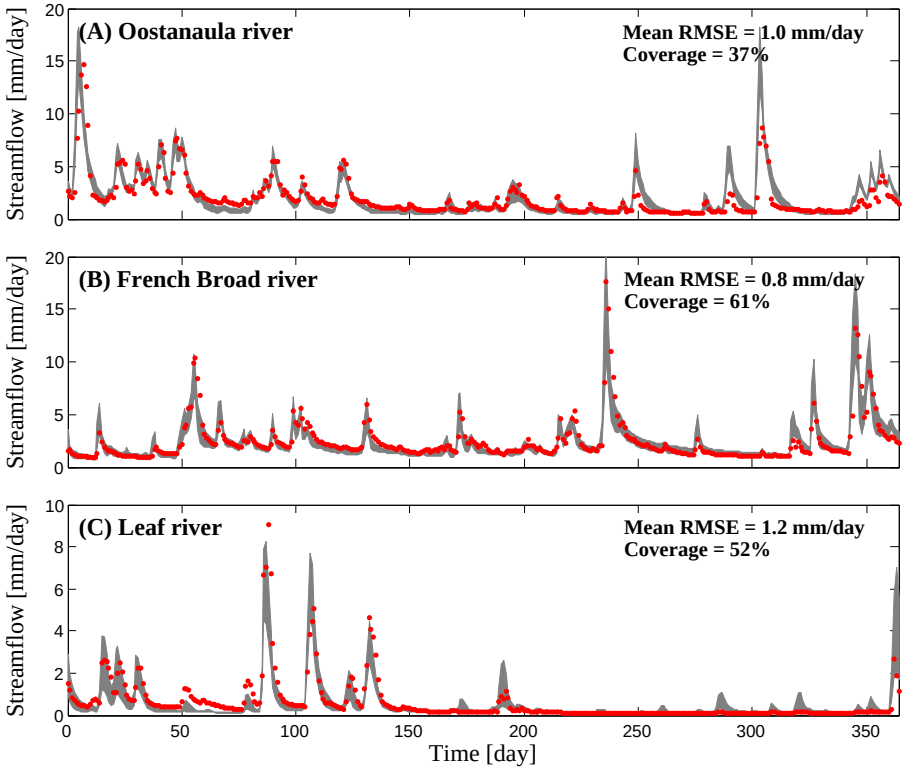


Figure 5.8: 95% posterior SAC-SMA simulation uncertainty ranges for the Oostanaula (a), French Broad (b) and Leaf (c) river watersheds for a representative 365 days portion of the calibration data period. The simulation uncertainty ranges (grey region) envelop between 37-61% of the streamflow observations (red dots).

catchment, and (b-c) two synthetic catchments. Each of the three Markov chains is coded with a different color. As a reminder, the first of these two synthetic catchments is referred to as "abrupt" and experiences a sudden change in the SAC-SMA model parameterization half way during the 10-year simulation period. The second artificial catchment is perhaps more realistic in that it experiences a gradual (linear) adjustment of the SAC-SMA parameter values during year 5-6 of the 10-year data period. This catchment is coined "gradual". The dashed red line depicts the value of $\epsilon = 0.05$ which is derived from the results in Table 5.1 and serves as benchmark to distinguish between stationary and nonstationary watersheds.

The sampled distance values in each of the Markov chains are (substantially) larger than the cutoff threshold of $\epsilon = 0.05$. In other words, the SAC-SMA model is unable to simulate

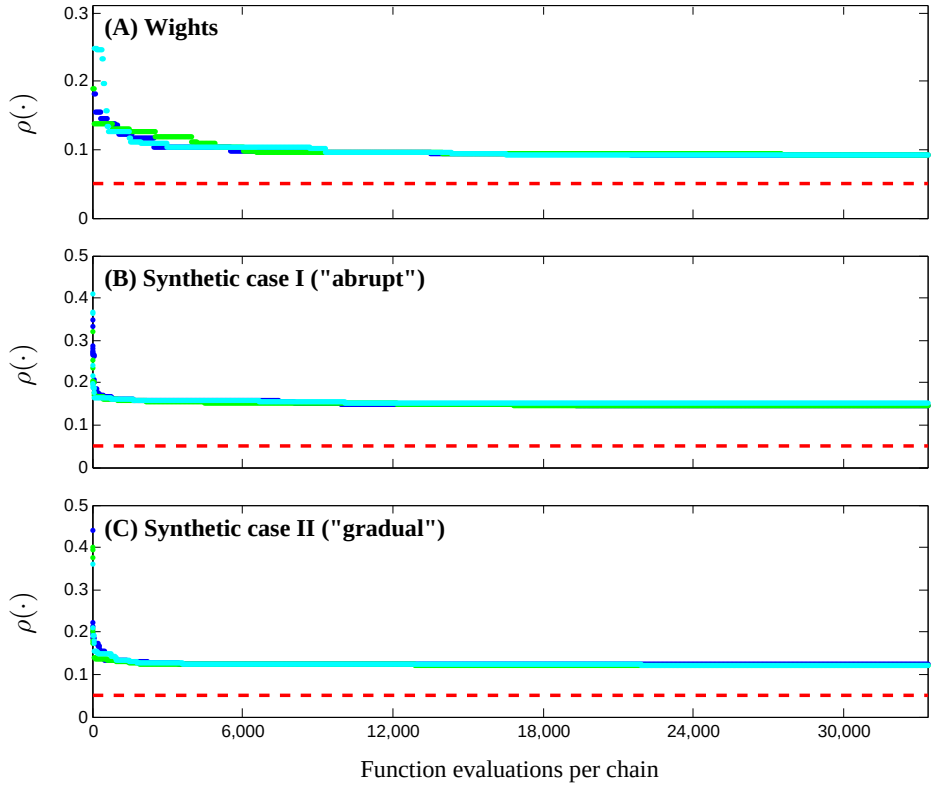


Figure 5.9: Trace plots of the sampled distance values, $\rho(\cdot)$, in each of the three Markov chains simulated with the DREAM$_{(\text{ABC})}$ algorithm for the Wights catchment (a), and the synthetic watersheds "abrupt" (b) and "gradual" (c). The dashed red line depicts the threshold of $\epsilon = 0.05$ used to distinguish between behavioral and nonbehavioral samples. The inability to find samples within epsilon of the measured summary metrics is interpreted as an incompatibility between the stationary assumption of the model structure and/or parameterization used, and the nonstationary behavior observed in the underlying data.

sufficiently close the observed values of each of the 14 summary metrics. Consequently, the samples derived from DREAM$_{(\text{ABC})}$ cannot be considered behavioral and should be discarded for posterior analysis. The sampled distance values are larger than the cutoff threshold of $\epsilon = 0.05$ because the SAC-SMA model is unable to adequately simulate the observed summary metrics with a single parameterization. These findings confirm our prior knowledge that the three watersheds have undergone changes. Note that the sampled value of $\rho(\cdot)$ are particularly large for the two synthetic watersheds. This is perhaps not surprising given the relatively large abrupt/gradual changes to each of the parameter values of the SAC-SMA model.

Note that the uncertainty of the parameters would be much smaller if a residual based likeli-

hood function is used [e.g. *Schoups and Vrugt* [2010a]]. However, many contributions to the hydrologic literature have established that such traditional statistical fitting method, albeit widely used, is unable to detect changes in the watersheds response to rainfall. Not only are the residuals very sensitive to errors in the precipitation data, but also they provide little guidance on model malfunctioning. The use of summary statistics has important diagnostic advantages, and combined with the ABC framework and sampling with DREAM$_{(\mathrm{ABC})}$ can help to elucidate whether the catchment has experienced changes or not. This is of particular importance to proclaim nonstationarity for the right reasons. Indeed, a single model parameterization is unable to mimic adequately temporal variations of the summary statistics (if any). Consequently, the simulated summary metrics deviate considerably from their observed counterparts and the values of $\rho(\cdot)$ from Equation 5.5 do not meet the benchmark of $\epsilon = 0.05$.

Finally, Figure 5.10 plots histograms of the SAC-SMA simulated summary statistics S_1, S_2, S_3, S_4 and S_{11} for each of the three nonstationary basins of Figure 5.9. These distributions are derived from the last 20% of the samples in each of the three Markov chains simulated with DREAM$_{(\mathrm{ABC})}$. The observed summary metrics are indicated separately in each plot with a red cross. The vertical dotted red lines in each panel denote the range of the behavioral (stationary) solution space.

Although some summary metrics are fitted sufficiently well with the SAC-SMA model, others appear too far removed from their observed counterparts to be classified as a behavioral solution. A particularly poor fit is observed for summary metric R_{11}, the RLD of year 7 of the data record. Perhaps this metric is the culprit and cannot be simulated adequately with a single model parameterization if the watershed has experienced urbanization, deforestation, and changes in land-use (amongst others) during the measurement data period. Thus, the RLD might exhibit important diagnostic power that can be used to detect changes in watershed behavior. Of course, a change in rainfall characteristics can affect the values of

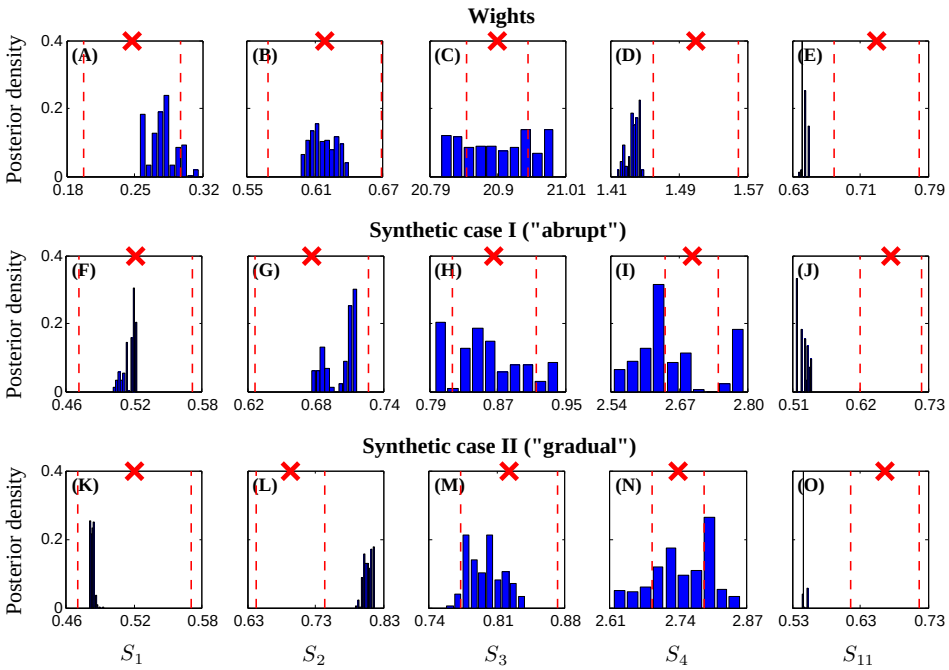


Figure 5.10: Histograms of the marginal posterior distributions of the runoff coefficient (S_1), baseflow index (S_2), FDC (S_3 and S_4) and rising limb density for the 7th year of the calibration period of the Wights catchment (top), and synthetic watersheds "gradual" (middle) and "abrupt" (bottom). The symbol "x" marks the measured values of the summary metrics. The "posterior" simulations of the SAC-SMA model cannot be considered behavioral as the summary metrics deviate too much from their observed counterparts. A single parameterization of the SAC-SMA model is unable to accurately describe each of the 14 observed summary metrics for each of the three basins. These results provide evidence that the watersheds behavior cannot be described adequately with a single model parameterization and hence has experienced a change during the calibration data period.

the summary statistics (e.g. a change to the FDC would lead to different values of its fitting coefficients) and suggest non-stationarity behavior. In other words, a value of $\epsilon > 0.05$ does not distinguish between changes in rainfall characteristics or catchment properties as cause for nonstationarity in the observed watershed response to precipitation. In any case, the methodology described herein provides hydrologists with a new approach to help determine whether watersheds exhibit stationary or nonstationary behavior. This concludes our numerical experiments.

A final remark is appropriate. In this work we have assumed a single value of ϵ that is used for each different summary metric. This approach is defensible in the present context as the summary metrics have a common scale and magnitude. Care should be exercised if new

metrics are used with disparate scales. Hence, it would seem logical to use a different ϵ for each individual summary metric. Temporal analysis of the values of the summary metrics can help to discern suitable values of ϵ for each individual metric. Alternatively, one could analyze the variability of the summary metrics for watersheds with similar climate, soil, and topographic properties. We leave this for future work. The summary metrics used in this study provide the diagnostic power to detect and pinpoint process changes in a watershed. Residual based likelihood functions, on the contrary, fail to provide such guidance.

5.5 Summary and conclusions

A key assumption in almost all modeling studies of the rainfall - runoff transformation is that the watershed properties are time-invariant and can be estimated from a historical record of discharge data emanating from the catchment outlet. This stationarity assumption has been questioned in several studies and publications in the past few years, in part due to evidence of an intensifying hydrologic cycle as well as the presence of trends and shift points in discharge dynamics of the watersheds response to rainfall, detected by statistical or other time series analysis methods. This behavior is classified as hydrologic nonstationarity, suggesting that the watershed has undergone changes in response to, for instance, urbanization and deforestation. Several authors cast doubt on the validity of such conclusions, in particular climate change and other long-term atmospheric disturbances can explain a large part of the observed trends and variations in annual and decadal streamflow dynamics. This long term persistency of the climate and weather is not readily apparent in short hydrologic data sets that span only 10-20 years. It would be highly desirable to have available a methodology that can determine whether the observed behavior is considered stationary or nonstationary. Ideally, such approach would also distinguish between the reasons of catchment nonstationarity, and determine whether the trends in streamflow dynamics are explained by decadal varia-

tions in weather and climate (rainfall), or whether the catchment properties have actually undergone changes as a result of human influences and alterations.

In this chapter, we have introduced a process-based model evaluation approach to help detect and diagnose the reasons of catchment nonstationarity. Our approach combines the strengths of diagnostic analysis, approximate Bayesian computation, and temporal analysis of watershed signatures (aka summary statistics) of the rainfall-runoff response to explain observed trends and/or shifts in discharge dynamics. Failure of a model to simulate accurately the observed temporal dynamics of the watersheds rainfall-runoff signatures is an indication of basin nonstationarity, and hence suggests that the catchment properties have changed during the time period under consideration. Indeed, a single model parameterization cannot adequately simulate temporal trends in the summary metrics of catchment response, and hence application of ABC cannot provide a behavioral model space.

Numerical simulations with artificial discharge data of the SAC-SMA model have identified the runoff coefficient, baseflow index, two fitting coefficients of the FDC, and annual rising limb density (RLD) to be the most suitable and informative summary metrics for diagnostic model evaluation, and analysis of nonstationarity. These metrics were selected from a large cohort of criteria and appear rather insensitive to rainfall data errors, an important prerequisite to avoid proclaiming nonstationarity based on temporally varying errors in the forcing data.

The analysis and findings presented in this chapter confirm our hypothesis that summary metrics of catchment behavior convey important information about hydrologic functioning and that temporal variations of these fingerprints is indicative of changes to the basins properties. Based on a suite of 19 different watersheds, 10 of which are deemed to exhibit stationary behavior, and 9 of which have been classified as nonstionary, we demonstrate that an ABC-threshold of $\epsilon = 0.05$ differentiates successfully between both groups of watersheds. If this threshold is applied to watersheds that are deemed nonstationary then the

DREAM$_{(\mathrm{ABC})}$ algorithm cannot provide a behavioral model space with simulations of the SAC-SMA model whose summary metrics are within ϵ of their observed counterparts.

The suggestion made in the hydrologic community that the Ferson and Blackberry creek basins in Australia exhibit nonstationary behavior is not confirmed by our diagnostic analysis with DREAM$_{(\mathrm{ABC})}$. The summary metrics of these watersheds appear rather constant over the duration of the 10-year data period, and the observed increases in streamflow volume of these two basins after 2006 is explained fully by a significant yearly increase in precipitation.

The annual rising limb density (RLD) appears most sensitive to changes in catchment behavior (properties), and exhibits a strong drift for nonstationary catchments. Thus, this summary metric contains important diagnostic power to help discern whether the catchment has experienced changes in its response to rainfall.

In summary, the methodology described herein provides hydrologists with a new approach to help diagnose and detect watershed nonstationarity. This framework is easy to use in practice and readily employs process-based metrics to capture different signatures of watershed behavior. Of course, additional testing of the proposed methodology against a much larger set of watersheds is required to further benchmark and refine our findings. We leave this for future work.

Chapter 6

The Soil Water Characteristic as New Class of Closed-form Parametric Expressions for the Flow Duration Curve

This Chapter has been submitted to Water Resources Research. Citation: Sadegh, M., J.A. Vrugt, and H.V. Gupta (2015), The soil water characteristic as new class of closed-form parametric expressions for the flow duration curve, *Water Resources Research*, Under Review.

6.1 Introduction

The flow duration curve (FDC) is a widely used characteristic signature of a watershed, and is one of the three most commonly used graphical methods in hydrologic studies, along with the mass curve and the hydrograph [*Foster*, 1934]. The FDC relates the exceedance probability (frequency) of streamflow to its magnitude, and characterizes both the flow regime and the streamflow variability of a watershed. It is closely related to "survival" function in statistics [*Vogel and Fennessey*, 1994], and is interpreted as a complement to the streamflow cumulative distribution function (CDF). The FDC is frequently used to predict the distribution of streamflow for water resources planning purposes, to simplify analysis of water resources problems, and to communicate watershed behavior to those who lack in-depth hydrologic knowledge. One should be particularly careful to rely solely on the FDC as main descriptor of catchment behavior [*Vogel and Fennessey*, 1995] as the curve represents the rainfall-runoff as disaggregated in the time domain and hence lacks temporal structure [*Searcy*, 1959, *Vogel and Fennessey*, 1994].

The first application of the FDC dates back to 1880 and appears in the work by Clemens Herschel [*Foster*, 1934]. Ever since, the FDC has been used in many fields of study including (among others) the design and operation of hydropower plants [*Singh et al.*, 2001], flow diversion and irrigation planning [*Chow*, 1964, *Warnick*, 1984, *Pitman*, 1993, *Mallory and McKenzie*, 1993], streamflow assessment and prediction [*Tharme*, 2003], sedimentation [*Vogel and Fennessey*, 1995], water quality management [*Mitchell*, 1957, *Searcy*, 1959, *Jehng-Jung and Bau*, 1996], waste-water treatment design [*Male and Ogawa*, 1984], and low-flow analysis [*Wilby et al.*, 1994, *Smakhtin*, 2001, *Pfannerstill et al.*, 2014]. Recent studies have used the FDC as a benchmark for quality control [*Cole et al.*, 2003], and signature or metric for model calibration and evaluation [*Refsgaard and Knudsen*, 1996, *Yu and Yang*, 2000, *Wagener and Wheater*, 2006, *Son and Sivapalan*, 2007, *Yadav et al.*, 2007, *Yilmaz et al.*, 2008, *Zhang et al.*, 2008, *Blazkova and Beven*, 2009, *Westerberg et al.*, 2011, *Vrugt and Sadegh*, 2013,

Pfannerstill et al., 2014, *Sadegh and Vrugt*, 2014, *Sadegh et al.*, 2015a]. For instance, *Vrugt and Sadegh* [2013] used the fitting coefficients of a simple parametric expression of the FDC as summary statistics in diagnostic model evaluation with approximate Bayesian computation (ABC).

Application of FDCs for hypothesis testing [*Kavetski et al.*, 2011] can improve identifiability and help attenuate the problems associated with traditional residual-based objective (likelihood) functions (e.g. Nash-Sutcliffe, sum of squared residuals, absolute error, relative error) that emphasize fitting specific parts of the hydrograph, such as high or low flows [*Schaefli and Gupta*, 2007, *Kavetski et al.*, 2011, *Westerberg et al.*, 2011], and thereby lose important information regarding the structural inadequacies of the model [*Gupta et al.*, 2008, 2012, *Vrugt and Sadegh*, 2013]. The FDC is a signature watershed characteristic that along with other hydrologic metrics, can help shed lights on epistemic (model structural) errors [*Euser et al.*, 2013, *Vrugt and Sadegh*, 2013]. For example, *Son and Sivapalan* [2007] used the FDC to highlight the reasons of model malfunctioning and to propose improvements to the structure of their conceptual water balance model for the watershed under investigation. Indeed, a deep groundwater flux was required to simulate adequately dominant low flows of the hydrograph. *Yilmaz et al.* [2008] in a similar effort to improve simulation of the vertical distribution of soil moisture in the HL-DHM model, used the slope of the FDC as benchmark for model performance. The FDC was deemed suitable for this purpose due to its strong dependence on the simulated soil moisture distribution, and relative lack of sensitivity to rainfall data and timing errors. However, the proposed refinements of the HL-DHM model were found inadequate and this failure was attributed to the inherent weaknesses of the conceptual structure of HL-DHM.

The usefulness of duration curves (e.g. precipitation [*Yokoo and Sivapalan*, 2011], baseflow [*Kunkle*, 1962] and streamflow (flow) [*Hughes and Smakhtin*, 1996]) depends in large part on the temporal resolution of the data (e.g. quarterly, hourly, daily, weekly, and monthly)

these curves are constructed from. FDCs derived from daily streamflow data are commonly considered to warrant an adequate analysis of the hydrologic response of a watershed [*Vogel and Fennessey*, 1994, *Smakhtin*, 2001, *Wagener and Wheater*, 2006, *Zhao et al.*, 2012]. For example, a FDC with a steep mid section (also referred to as slope) is characteristic for a watershed that responds quickly to rainfall, and thus has a small storage capacity and large ratio of direct runoff to baseflow. A more moderate slope, on the contrary, is indicative of a basin whose streamflow response reacts much slower to precipitation forcing with discharge that is made up in large part of baseflow [*Yilmaz et al.*, 2008].

The shape of the FDC is determined by several factors including (amongst others) topography, physiography, climate, vegetation cover, land use, and storage capacity [*Singh*, 1971, *Lane et al.*, 2005, *Zhao et al.*, 2012], and can be used to perform regional analysis [*Wagener and Wheater*, 2006, *Masih et al.*, 2010] or to cluster catchments into relatively homogeneous groups that exhibit a relatively similar hydrologic behavior [*Coopersmith et al.*, 2012, *Sawicz et al.*, 2011]. Different studies have appeared in the hydrologic literature that have analyzed how the shape of the FDC is affected by physiographic factors and/or vegetation cover. Despite this progress made, interpretation of the FDC can be controversial if an insufficiently long streamflow data record is used [*Vogel and Fennessey*, 1994]. The lower end of the FDC (low flows) is particularly sensitive to the period of study, and to whether the streamflow data includes severe droughts or not [*Castellarin et al.*, 2004a]. If the available data is spare and does not warrant an accurate description of the FDC, then the use of an annual duration curve is advocated [*Searcy*, 1959, *Vogel and Fennessey*, 1994, *Castellarin et al.*, 2004a,b]. This curve describes the relationship between the magnitude and frequency of the streamflow for a "typical hypothetical year" [*Vogel and Fennessey*, 1994]. To construct an annual FDC, the available data is divided into n years and individual FDCs are constructed for each year. Then, for each exceedance probability a median streamflow is derived from these n different FDCs and used to create the annual FDC. *Vogel and Fennessey* [1994] used this concept to associate confidence and recurrence intervals to FDCs in a nonparametric framework. One

should note that the FDC of the total data record is, in general, more accurate than the annual FDC [*Leboutillier and Waylen*, 1993].

To better analyze and understand the physical controls on the FDC, it is common practice to divide the total FDC (TFDC) into a slow (SFDC) and fast (FFDC) flow component [*Yokoo and Sivapalan*, 2011, *Cheng et al.*, 2012, *Coopersmith et al.*, 2012, *Yaeger et al.*, 2012, *Ye et al.*, 2012]. For example, *Yokoo and Sivapalan* [2011] concluded from numerical simulations with a simple water balance model that the FFDC is controlled mainly by precipitation events and timing, whereas the SFDC is most sensitive to the storage capacity of the watershed and its baseflow response. This type of analysis is of particular value in regionalization studies, and prediction in ungauged basins. Indeed, much effort has gone towards prediction of the FDC in ungauged basins using measurements of the rainfall-runoff response from hydrologically similar, and preferably geographically nearby, gauged basins [*Holmes et al.*, 2002, *Sivapalan et al.*, 2003].

In this context, one approach has been to cluster catchments into classes with similar physiographic and climatic characteristics, and then to estimate dimensionless (non-parametric) FDCs for gauged basins which in turn are then applied to ungauged basins [*Niadas*, 2005, *Ganora et al.*, 2009]. The FDCs are normalized by an index value (e.g. mean annual runoff) to generate dimensionless curves [*Ganora et al.*, 2009]. A detailed review on methods for clustering of homogeneous catchments appears in *Sauquet and Catalogne* [2011] and *Booker and Snelder* [2012], and interested readers are referred to these publications for more information. Another approach has been to mimic the empirical (observed) FDC with a mathematical/probabilistic model and to correlate the fitting coefficients of such parametric expressions to physical and climatological characteristics of the watershed using regression techniques [*Fennessey and Vogel*, 1990, *Yu et al.*, 1996, 2002, *Croker et al.*, 2003, *Castellarin et al.*, 2004a,b, 2007, *Li et al.*, 2010, *Sauquet and Catalogne*, 2011]. Such pedotransfer functions can then be used to predict the FDC of ungauged basins from simple catchment data

(e.g. soil texture, topography, vegetation cover, etc.).

Models that emulate the FDC can be grouped in two main classes: 1. Physical models that use physiographic and climatic characteristics of the watersheds (e.g. drainage area, mean areal precipitation, soil properties, etc.) as parameters of the FDC [*Singh*, 1971, *Dingman*, 1978, *Yu et al.*, 1996, *Holmes et al.*, 2002, *Yu et al.*, 2002, *Lane et al.*, 2005, *Botter et al.*, 2008, *Mohamoud*, 2008]; and, 2. Probabilistic/mathematical functions that use between two to five fitting coefficients to mimic the empirical FDC as closely and consistently as possible [*Quimpo et al.*, 1983, *Mimikou and Kaemaki*, 1985, *Fennessey and Vogel*, 1990, *Leboutillier and Waylen*, 1993, *Franchini and Suppo*, 1996, *Cigizoglu and Bayazit*, 2000, *Croker et al.*, 2003, *Botter et al.*, 2008, *Li et al.*, 2010, *Booker and Snelder*, 2012].

The early work of *Dingman* [1978] is the first study that used physical models to mimic empirical FDCs. Topography maps were used as signature of catchment behavior to predict the FDC using relatively simple first-order polynomial functions. Two more recent studies by *Yu et al.* [1996, 2002] used similar regression functions to predict the FDC of catchments in Taiwan but considered the drainage area as main proxy of the rainfall-runoff transformation.

Probabilistic methods include the use of lognormal and lognormal mixture [*Fennessey and Vogel*, 1990, *Leboutillier and Waylen*, 1993, *Castellarin et al.*, 2004a, *Li et al.*, 2010], generalized Pareto [*Castellarin et al.*, 2004b], generalized extreme value, gamma and Gumbel [*Booker and Snelder*, 2012], beta [*Iacobellis*, 2008], and logistic distributions [*Castellarin et al.*, 2004b]. Unfortunately, the presence of temporal correlation between successive (e.g. daily) streamflow observations violates some of the basic assumptions of these classical statistical distributions, hence casting doubt on the validity of such models to describe closely and consistently the FDC [*Vogel and Fennessey*, 1994]. Other mathematical models include (amongst others) the use of exponential, power, logarithmic [*Quimpo et al.*, 1983, *Franchini and Suppo*, 1996, *Lane et al.*, 2005, *Booker and Snelder*, 2012], and polynomial functions [*Mimikou and Kaemaki*, 1985, *Yu et al.*, 2002]. In another line of work, *Cigizoglu and Bayazit*

[2000] used convolution theory to predict the FDC as a product of periodic and stochastic streamflow components. What is more, *Croker et al.* [2003] used probability theory to combine a model that predicts the FDC of days with non-zero streamflow with a distribution function that determines randomly the probability of dry days.

Notwithstanding this progress made, the probabilistic and mathematical functions used in the hydrologic literature fail, usually, to properly mimic all parts of the FDC when benchmarked against watersheds with completely different hydrologic behaviors. Consequently, some researchers focus only on a specific portion of the FDC, commonly the low flows [*Fennessey and Vogel*, 1990, *Franchini and Suppo*, 1996], whereas others prefer to use several percentiles of the FDC rather than the entire curve [*Lane et al.*, 2005, *Mohamoud*, 2008, *Blazkova and Beven*, 2009]. While complex "S" shaped FDCs can only be modeled adequately if a sufficient number of parameters (say five to seven) are used [*Ganora et al.*, 2009], one should be particularly careful using such relatively complex FDC models for regionalization and prediction in ungauged basins as parameter correlation and insensitivity can complicate and corrupt the inference and results.

In this chapter, we introduce a new class of closed-form mathematical expressions which are capable of describing the FDCs of a very large number of watersheds with contrasting hydrologic behaviors. We extend the ideas presented in *Vrugt and Sadegh* [2013], and propose several commonly used functions of the soil water characteristic as mathematical models for the observed (empirical) FDCs. These models have between two to five parameters and describe closely the FDCs of the MOPEX data set. In fact, a watershed can be viewed as a large porous medium that holds and transports water depending on the soil moisture content and applied boundary conditions (rainfall and potential evapotranpiration). This analogy opens up many new possibilities for analyzing the FDC and provides us with an arsenal of new mathematical expressions that can be used for (amongst others) hydrologic modeling, prediction in ungauged basins and regionalization.

This work is a follow up of Chapter 2 in which we introduced a 2-parameter formulation of the van Genuchten (VG) retention function [*van Genuchten*, 1980] as model for the FDC. The fitting coefficients of this function were derived by calibration against the observed FDC of the French Broad river basin and used as summary statistics in diagnostic model evaluation with ABC. Here, we extend this preliminary work, and introduce and benchmark several other soil water retention functions (SWRFs) using historical streamflow observations from 408 watersheds of the MOPEX data set. We are particularly interested in the quality of fit of each model, parameter sensitivity and correlation, and regionalization potential of the fitting coefficients.

The remainder of this chapter is organized as follows. Section 6.2 provides a brief review of the most commonly used probabilistic and empirical models in the literature to mimic the FDC, and introduces several closed-form expressions of the soil water characteristic to emulate FDCs. This section also discusses the MOPEX data set, and the optimization procedure to estimate the values of the fitting coefficients of each SWRF. Then in section 6.3, we present and discuss the fitting results of each of the proposed parametric expressions. Here, we are especially concerned with benchmarking of our results against those of the FDC models used in the literature, analysis of parameter sensitivity and correlation, and regionalization potential of the fitting coefficients of the SWRFs. Finally, Section 6.4 concludes this chapter with a summary of the most important findings.

6.2 Materials and methods

6.2.1 Experimental data: MOPEX data set

We use daily streamflow data from 438 MOPEX watersheds to analyze the ability of different mathematical functions to fit the empirical FDCs. This includes watersheds from across the

United States with different hydrologic behaviors. The empirical FDCs are constructed using all of the available streamflow data for each watershed. Figure 6.1 shows the observed FDCs for eight of the MOPEX watersheds. The y-axis is scaled to a maximum value of one so that the FDCs of the watersheds can be compared more easily visually. A large variety in the FDCs of the MOPEX data set is observed. This confirms the need for a flexible function that can adequately describe the large range of hydrologic behaviors.

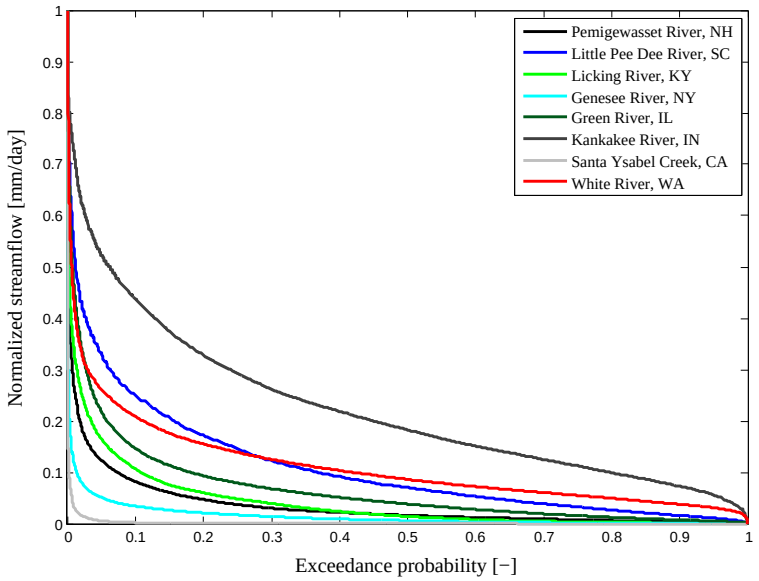


Figure 6.1: Flow duration curves for eight watersheds from the MOPEX data set, which show contrasting hydrologic behaviors. This figure plots normalized streamflows (mm/day) versus exceedance probability (-). We normalized the streamflow values to plot various shapes of the FDCs in one figure.

This rich archive of 438 catchments in the USA contains relatively long records of daily hydrologic data (rainfall, potential evapotranspiratin and streamflow). Six watersheds were excluded from our analysis due to a lack of streamflow observations, and another 24 catchments were removed due to the presence of a very large number of days with zero flow. The remaining 408 catchments have between 236 and 19,997 daily streamflow observations and were used for FDC model testing and evaluation. Obviously, a 236 day record is insufficient to warrant an accurate description of the FDC - nevertheless this data length is still appropriate for testing of the proposed models. Note, for 213 watersheds more than 54 years of

daily streamflow data is available.

6.2.2 Probabilistic and mathematical functions

The functions most widely used in the literature to fit the observed FDCs are briefly reviewed. The skew and peakiness so clearly visible in histograms of long data records of daily streamflow prompts the use of the Gumbel distribution to model empirical FDCs [*Booker and Snelder*, 2012]. The CDF of the Gumbel distribution is given by

$$\mathcal{F}_{\mathbf{Y}}(y_t|a_{\mathrm{G}}, b_{\mathrm{G}}) = \exp\left[-\exp\left(-\frac{y_t - a_{\mathrm{G}}}{b_{\mathrm{G}}}\right)\right], \tag{6.1}$$

where $\mathcal{F}_{\mathbf{X}}$ denotes the cumulative density, a_{G} signifies the location, and b_{G} measures the scale of y_t. The Gumbel distribution is a particular case of the generalized extreme value (GEV) distribution, whose CDF is given by

$$\mathcal{F}_{\mathbf{Y}}(y_t|a_{\mathrm{GEV}}, b_{\mathrm{GEV}}, c_{\mathrm{GEV}}) = \exp\left(-\left[1 + c_{\mathrm{GEV}}\left(\frac{y_t - a_{\mathrm{GEV}}}{b_{\mathrm{GEV}}}\right)\right]^{-1/c_{\mathrm{GEV}}}\right), \tag{6.2}$$

where a_{GEV}, b_{GEV}, and c_{GEV} are the location, scale, and shape parameters, respectively. The shape parameter, c_{GEV}, controls the skewness of the distribution, and enables fitting of tailed streamflow distributions. The GEV distribution is widely used in the field of hydrology to model extremes (floods and droughts) [*Katz et al.*, 2002], as well as FDCs [*Booker and Snelder*, 2012]. Note that if $c_{\mathrm{GEV}} = 0$, the GEV distribution simplifies to the Gumbel distribution. The exceedance probability, G_t, of a streamflow observation/simulation, y_t, can be derived from Equations 6.1 and 6.2 by simply calculating

$$G_t = 1 - \mathcal{F}_{\mathbf{Y}}(y_t|\cdot). \tag{6.3}$$

This concludes our description of the probabilistic models of the FDC. We continue our review with two widely used mathematical functions of the FDC, one of which is the 2-parameter exponential function of *Quimpo et al.* [1983] which is given by

$$y_t = a_{\mathrm{Q}} \exp(-b_{\mathrm{Q}} G_t), \tag{6.4}$$

where a_{Q} (mm/day) and b_{Q} (-) are fitting coefficients. This function predicts streamflow, y_t, based on exceedance probability, G_t, while the CDFs predict exceedance probability, G_t, based on the observed discharge, y_t. We can simply invert Equation 6.4 to derive a mathematical expression for the exceedance probability, G_t, whose value can be compared directly against the results of Equation 6.3 for the probabilistic models

$$G_t = -\frac{1}{b_{\mathrm{Q}}} \log\left(\frac{y_t}{a_{\mathrm{Q}}}\right). \tag{6.5}$$

Franchini and Suppo [1996] proposed another function to model empirical FDCs. Their 3-parameter model is defined as

$$y_t = b_{\mathrm{FS}} + a_{\mathrm{FS}}(1 - G_t)^{c_{\mathrm{FS}}}, \tag{6.6}$$

in which a_{FS} (mm/day), b_{FS} (mm/day) and c_{FS} (-) are fitting coefficients. This function was originally proposed to only fit the low flows of the FDC, but has been used to model the entire range of the FDC as well [*Sauquet and Catalogne*, 2011]. Inversion of Equation 6.6 provides a direct expression for the exceedance probability

$$G_t = 1 - \left(\frac{y_t - b_{\mathrm{FS}}}{a_{\mathrm{FS}}}\right)^{1/c_{\mathrm{FS}}}, \tag{6.7}$$

so that we can compare this model against the CDF of the probabilistic models.

Practical experience suggests that the parametric models reviewed thus far are not flexible

enough to warrant a close description of the FDC for all different exceedance probabilities, let alone capture adequately the large variability that exists among the FDCs of different watersheds. In the next section we introduce a new class of closed-form parametric expressions of the FDC which accurately describe the FDCs for a large range of hydrologic behaviors.

6.2.3 Proposed formulations

The functional shape of the FDC has many elements in common with that of the soil water characteristic. This is graphically illustrated in Figure 6.2 which plots the water retention function of five different soils presented in *van Genuchten* [1980]. These curves depict the relationship between the volumetric moisture content, θ (x-axis) and the corresponding pressure head, h (y-axis) of a soil and are derived by fitting Equation

$$\theta = \theta_r + \frac{\theta_s - \theta_r}{[1 + (\alpha|h|)^n]^m}, \tag{6.8}$$

to experimental (θ, h) data collected in the laboratory. This equation is also known as the van Genuchten (VG) model and contains five different parameters, where θ_s ($\mathrm{cm}^3/\mathrm{cm}^3$) and θ_r ($\mathrm{cm}^3/\mathrm{cm}^3$) denote the saturated and residual moisture content, respectively, and α (1/cm), n (-) and m (-) are fitting coefficients that determine the air-entry value and slope of the WRF. In most studies, the value of m is set conveniently to $1 - 1/n$ which not only reduces the number of parameters to four, but also provides a closed-form expression for the unsaturated soil hydraulic conductivity function [*van Genuchten*, 1980].

The shape of the WRFs plotted in Figure 6.2 show great similarity with the FDCs displayed previously in Figure 6.1. This suggests that Equation 6.8 might be an appropriate parametric expression for describing the functional relationship between the exceedance probability of streamflow (x-axis) and its magnitude (y-axis). To make sure that the exceedance probability is bounded exactly between 0 and 1, we set $\theta_r = 0$ and $\theta_s = 1$, respectively. This leads to

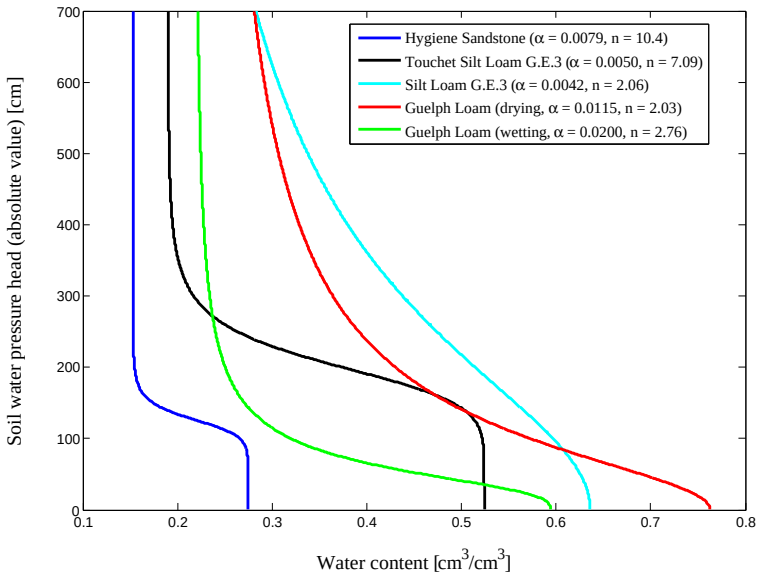


Figure 6.2: Water retention curves for five different soil types introduced in *van Genuchten* [1980]. This figure plots soil water pressure head (cm) versus soil water content (cm^3/cm^3). Note that the soil water pressure head values are originally negative, but we use positive pressure heads for plotting since the absolute values are used in the WRFs.

the following 2-parameter VG formulation of the FDC proposed by *Vrugt and Sadegh* [2013]

$$G_t = \left[1 + (a_{\mathrm{VG}} y_t)^{b_{\mathrm{VG}}}\right]^{1/b_{\mathrm{VG}}-1}, \tag{6.9}$$

where the parameters a_{VG} (day/mm), and b_{VG} (-) need to determined by fitting against the observed FDC. If deemed appropriate, we can further increase the flexibility of Equation 6.9 by defining $c_{\mathrm{VG}} = 1/b_{\mathrm{VG}} - 1$. In this chapter, we will test both the 2- and 3-parameter VG formulations of the FDC.

The VG model is used widely in porous flow simulators to describe numerically variably saturated water flow. Yet, many other hydraulic models have been proposed in the vadose zone literature to characterize the retention and unsaturated soil hydraulic conductivity

functions. We consider here the lognormal WRF of *Kosugi* [1994, 1996]

$$G_t = \begin{cases} \frac{1}{2}\,\mathrm{erfc}\left[\frac{1}{\sqrt{2}b_{\mathrm{K}}}\log\left(\frac{y_t - c_{\mathrm{K}}}{a_{\mathrm{K}} - c_{\mathrm{K}}}\right)\right] & \text{if } y_t > c_{\mathrm{K}} \\ \\ 1 & \text{if } y_t \leq c_{\mathrm{K}} \end{cases}, \tag{6.10}$$

where "erfc" denotes the complimentary error function, and a_{K} (mm/day), b_{K} (-) and c_{K} (mm/day) are fitting coefficients that need to be determined by calibration against the empirical FDC of each watershed. We can simplify this 3-parameter Kosugi formulation of the FDC by setting $c_{\mathrm{K}} = 0$. This assumption is appropriate for experimental data of the soil water characteristic, and we therefore consider both the 2- and 3-parameter formulations in our numerical studies with the MOPEX data set presented later.

The main reason to use Kosugi's WRF rather than other commonly used parametric expressions of the WRF such as *Brooks and Corey* [1964], is that the parameters can be related directly to the pore size distribution and hence exhibit a much better physical underpinning than their counterparts of the VG model. This might increase the chances of regionalization efforts.

The 2- and 3-parameter formulations of the Kosugi and VG models of the WRF assume implicitly a unimodal pore size distribution. This assumption might not be realistic for some soils and can lead to a relatively bad fit of these unimodal pore size distribution models to the observed water retention data. Soils with a heterogeneous (e.g. bimodal) pore size distribution can be much better described if we use a mixture of two or more WRFs. We follow *Durner* [1994] and propose the following mixture FDCs for watersheds with a

complex hydrologic response

$$
\begin{aligned}
G_t &= \sum_{i=1}^{k} \left(w_{i,\mathrm{VG}} \left[1 + (a_{i,\mathrm{VG}}\, y_t)^{b_{i,\mathrm{VG}}} \right]^{1/b_{i,\mathrm{VG}}-1} \right) \\
G_t &= \sum_{i=1}^{k} \left(\frac{w_{i,\mathrm{K}}}{2} \operatorname{erfc} \left[\frac{1}{\sqrt{2} b_{i,\mathrm{K}}} \log \left(\frac{y_t}{a_{i,\mathrm{K}}} \right) \right] \right),
\end{aligned}
\tag{6.11}
$$

where k denotes the number of constituent modes, $w_{i,\mathrm{X}}$ signifies the weight of each individual WRF ($w_{i,\mathrm{X}} \geq 0$ and $\sum_{i=1}^{k} w_{i,\mathrm{X}} = 1$), and $a_{i,\mathrm{X}}$ and $b_{i,\mathrm{X}}$ are the fitting coefficients of the ith WRF. The subscript X stands either for VG or K. If fitting of the FDC is of main importance, then one could use a relatively large value for k (e.g. $k = 5$). Yet, for diagnostic analysis and regionalization one should be particularly careful not to overfit the empirical FDCs. Indeed, considerable trade-off exists between the quality of fit of the FDC and the identifiability and correlation of the FDC model coefficients. The more parameters that are used to emulate the FDC, the better the fit to the empirical curve but at the expense of an increase in parameter uncertainty and correlation. In the present study we use $k = 2$ and hence mimic the empirical FDCs with a mixture of two VG or two Kosugi WRFs. For both classes of models, this mixture formulation has five parameters.

A bimodal FDC would seem appropriate for catchments with two or more discerning flow paths. For example, if the catchment exhibits two dominant modes (e.g. surface runoff and baseflow) then a bimodal mixture formulation of the FDC would seem appropriate. Table 6.1 summarizes the 2-, 3- and 5-parameter formulations of VG and Kosugi used herein to model the FDC. These models will be tested against the empirical FDCs of the MOPEX data set. These results will be presented in the next section. For completeness we have also tabulated the probabilistic models used herein.

Table 6.1: Summary of the probablistic, mathematical and WRF-based functions of the FDC. All these functions compute the exceedance probability, G_t for a given streamflow value, y_t. The variables a, b, and c are fitting coefficients whose values need to be derived by calibration against the observed, empirical, FDC of individual watersheds.

Model name	Model formulation
	2-parameter formulations
Gumbel	$G_t = 1 - \exp\left[-\exp\left(-\frac{y_t - a_{\mathrm{G}}}{b_{\mathrm{G}}}\right)\right]$
Quimpo	$G_t = -\frac{1}{b_{\mathrm{Q}}} \log\left(\frac{y_t}{a_{\mathrm{Q}}}\right)$
VG	$G_t = \left[1 + (a_{\mathrm{VG}} y_t)^{b_{\mathrm{VG}}}\right]^{1/b_{\mathrm{VG}} - 1}$
Kosugi	$G_t = \frac{1}{2} \operatorname{erfc}\left[\frac{1}{\sqrt{2} b_{\mathrm{K}}} \log\left(\frac{y_t}{a_{\mathrm{K}}}\right)\right]$
	3-parameter formulations
GEV	$G_t = 1 - \exp\left(-\left[1 + c_{\mathrm{GEV}}\left(\frac{y_t - a_{\mathrm{GEV}}}{b_{\mathrm{GEV}}}\right)\right]^{-1/c_{\mathrm{GEV}}}\right)$
Franchini and Suppo	$G_t = 1 - \left(\frac{y_t - b_{\mathrm{FS}}}{a_{\mathrm{FS}}}\right)^{1/c_{\mathrm{FS}}}$
VG	$G_t = \left[1 + (a_{\mathrm{VG}} y_t)^{b_{\mathrm{VG}}}\right]^{-c_{\mathrm{VG}}}$
Kosugi	$G_t = \begin{cases} \frac{1}{2} \operatorname{erfc}\left[\frac{1}{\sqrt{2} b_{\mathrm{K}}} \log\left(\frac{y_t - c_{\mathrm{K}}}{a_{\mathrm{K}} - c_{\mathrm{K}}}\right)\right] & \text{if } y_t > c_{\mathrm{K}} \\ 1 & \text{if } y_t \leq c_{\mathrm{K}} \end{cases}$
	5-parameter formulations
VG	$G_t = \sum_{i=1}^{2} \left(w_{i,\mathrm{VG}}\left[1 + (a_{i,\mathrm{VG}} y_t)^{b_{i,\mathrm{VG}}}\right]^{1/b_{i,\mathrm{VG}} - 1}\right)$
Kosugi	$G_t = \sum_{i=1}^{2} \left(\frac{w_{i,\mathrm{K}}}{2} \operatorname{erfc}\left[\frac{1}{\sqrt{2} b_{i,\mathrm{K}}} \log\left(\frac{y_t}{a_{i,\mathrm{K}}}\right)\right]\right)$

6.2.4 FDC model parameter estimation

Now we have discussed the different models for the FDC (see Table 6.1), we are left with inference of their coefficients. We have developed a MATLAB program called FDC_Fitting which automatically determines the best values of the fitting coefficients of the FDC mod-

els used herein for a given streamflow data record. Graphical output is provided as well. FDC_Fitting implements a multi-start gradient-based local optimization approach with Levenberg Marquardt (LM) [*Marquardt*, 1963] to rapidly locate the "best" values of the FDC model parameters. A classical sum of squared error (SSE) objective function is used to summarize the distance between the observed and simulated FDCs,

$$\mathrm{SSE}(\mathbf{x}|\tilde{\mathbf{Y}}) = \sum_{t=1}^{N} \left(\tilde{G}_t - G_t(\mathbf{x}|\tilde{y}_t)\right)^2, \tag{6.12}$$

where $\mathbf{x}$ is the vector of fitting parameters (differs per FDC model), $\tilde{\mathbf{Y}} = \{\tilde{y}_1, \ldots, \tilde{y}_N\}$ denotes the observed streamflow data, and $\tilde{G}_t$ and G_t signify the observed and FDC modeled exceedance probabilities, respectively. To maximize LM search efficiency, the Jacobian matrix is computed using analytical expressions for $\partial G_t/\partial x_i$ ($i = \{1, \ldots, d\}$), the partial sensitivity of the output (exceedance probability) predicted by each FDC model with respect to its individual model parameters. Moreover, 20 different LM trials were used each with a different starting point, drawn randomly from $\mathcal{U}_d[0, 1]$, the d-dimensional uniform distribution on the interval between zero and one.

The LM search method is computationally efficient and much faster than global optimization methods. What is more, it does not require specification of a prior parameter space (except that all parameters are at least zero). Nevertheless, in some cases ($< 5\%$) the LM method was unable to reduce substantially the fitting errors of the (randomly chosen) starting points and the (derivative-free) Nelder-Mead Simplex algorithm [*Nelder and Mead*, 1965] was used instead to minimize Equation 13. Numerical results for the 3- and 5-parameter VG and Kosugi FDC models demonstrate (not shown) that this hybrid two-step approach provides better results for some of the watersheds than commonly used global optimizers. This is in large part due to the rather peculiar properties of the SSE-based response surfaces (flat with one small solution pocket).

To derive explicit estimates of FDC parameter uncertainty, we also solved for the posterior distribution of the fitting coefficients using Bayesian inference with DREAM [*Vrugt et al.*, 2008b, 2009a]. We assume a uniform prior distribution for all the FDC parameters, and sample the target distribution by maximizing a standard Gaussian (least-squares type) log-likelihood function, $\log\big(p(\mathbf{x}|\tilde{\mathbf{Y}})\big) = -N/2\log\big(\mathrm{SSE}(\mathbf{x}|\tilde{\mathbf{Y}})\big)$ where p denotes the posterior density. In our application of DREAM, we used default settings of the algorithmic variables, and convergence of the parallel chains to a limiting distribution was assessed with different diagnostics.

6.3 Results and discussion

Table 6.2 summarizes the performance of each of the FDC models for the MOPEX data set used herein. We list separately the first (mean) and second-order (standard deviation) moment of the Root Mean Square Error (RMSE) of the fit to the empirical FDCs of the 408 different watersheds. For model selection purposes we also list the number of times (expressed in %) each respective FDC model with two, three or five parameters achieves the lowest RMSE among its constituent formulations with similar number of parameters (degrees of freedom).

The results in Table 6.2 highlight several key findings. Firstly, the 2-parameter FDC formulations of Kosugi and VG exhibit substantially lower mean RMSE values than their counterparts with similar structural complexity (Gumbel, Quimpo, and Franchini and Suppo) proposed in the literature. Secondly, of the four different 2-parameter models, the Kosugi WRF is preferred for more than 60% of the watersheds as this model best fits the empirical FDCs. What is more, the two parameters of this parsimonious formulation of Kosugi exhibit a negligible correlation (shown later) which constitutes an important practical advantage for model fitting, regionalization and diagnostic model evaluation (discussed later). Thirdly,

among the 3-parameter formulations of the FDC, the Kosugi, VG and GEV models exhibit a somewhat similar performance (according to RMSE), and their fit is much better than that of the model of Franchini and Suppo. Nevertheless, the Kosugi WRF is most supported by the available data as it achieves the lowest RMSE for about 56% of the catchments. Finally, the 5-parameter formulations of Kosugi and VG best fit the empirical FDCs of the MOPEX data set. The mean RMSE of these two models ranges between 0.008 - 0.009 (computed in exceedance space) and is substantially lower than values of 0.024 - 0.029 and 0.016 - 0.017 for their 2- and 3-parameter variants, respectively. Altogether, our results favor the Kosugi WRF as model for the FDC as this WRF consistently provides the closest fit to the observed FDCs (in terms of mean RMSE) for all different model complexities.

Table 6.2: Performance statistics of the quality of fit of each individual FDC model. We list the mean and standard deviation of the Root Mean Square Error (RMSE) (derived from Equation 6.12), and the number of times (in %) each respective model within a certain parametric class provides the best fit (lowest RMSE) to the observed FDCs.

Model name	Mean RMSE	STD of RMSE	% Best model
2-parameter formulations			
Gumbel	0.093	0.016	0
Quimpo	0.066	0.026	0.7
VG	0.026	0.012	39.2
Kosugi	0.021	0.014	60.1
3-parameter formulations			
GEV	0.015	0.008	9.8
Franchini and Suppo	0.081	0.021	0.3
VG	0.014	0.007	30.6
Kosugi	0.013	0.008	59.3
5-parameter formulations			
VG	0.006	0.003	30.9
Kosugi	0.005	0.003	69.1

To provide more insights into the fitting results for individual catchments, consider Figure 6.3 that plots histograms of the RMSE values of the 408 different watersheds. The left

column (A-D) plots the results for the 2-parameter FDC models, whereas the middle and right columns illustrate the results for the 3- and 5-parameter models, respectively. Color coding is used to differentiate among the FDC models. For instance, the color black in plot A, E and I refers to VG, blue in B, G, and J signifies the Kosugi model, and the other colors (cyan, light and dark green, and gray) are used for the remaining models proposed in the literature. To simplify visual interpretation a common x-axis was used for all different models and complexities.

The results presented in Figure 6.3 confirm our earlier conclusions that the proposed WRF-based FDC models are superior to the probabilistic and mathematical FDC models proposed and used in the literature. For the 2-parameter functions (left column, plot B), the Kosugi model is preferred as its distribution of RMSE values is most oriented towards the left in direction of zero RMSE. The distributions of the RMSE values of the 3-parameter formulations of VG and Kosugi (plot E and F) are quite similar, and in close agreement with the histogram derived from the GEV model (plot G). These three models perform substantially better than the model of Franchini and Suppo (plot H), whose distribution of RMSE values is approximately Gaussian but with a much larger mode (about 0.08) and standard deviation. The mixture formulations of VG and Kosugi with five parameters respectively, exhibit the best overall performance. The mode of their RMSE histogram is in closest vicinity of zero and the dispersion (standard deviation) of this distribution is smallest of all different model complexities. Overall, the 5-parameter Kosugi function provides the most accurate fit to the empirical FDCs of the 408 MOPEX watersheds, and should be used if the main purpose for application of these models is solely to predict the exceedance probability for a given streamflow value.

Thus far we have focused our attention on the quality of fit of the VG and Kosugi FDC models without recourse to assessment of their parameter sensitivity, identifiability and correlation. Figure 6.4 plots, for a large range of streamflow values, the sensitivity of the simulated ex-

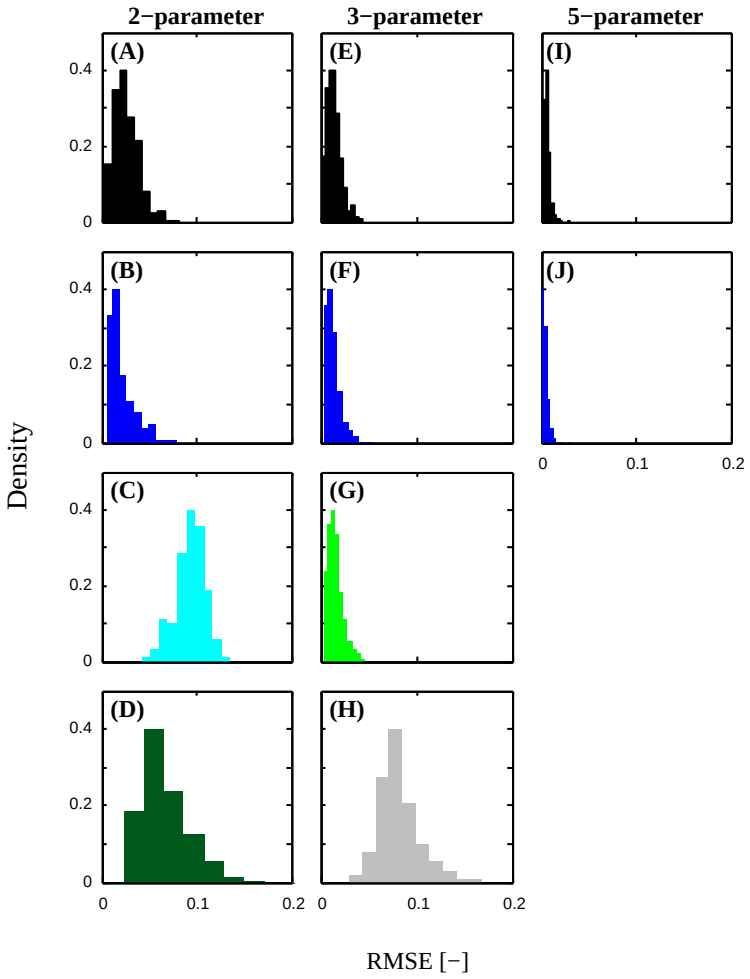


Figure 6.3: Histogram of the root mean square errors (RMSEs) between the simulated and observed FDCs for the 408 studied watersheds. The x-axis in each plot represents the RMSE values (dimensionless) and the y-axis shows the density of the number of occurrences for each RMSE value. Left, middle and right columns show the results of 2-, 3-, and 5-parameter model formulations, respectively. The results of the VG (plots: A, E, I), Kosugi (plots: B, F, J), Gumbel (plot: C), Quimpo (plot: D), GEV (plot: G), and Franchini and Suppo (plot: H) are presented with black, blue, cyan, dark green, light green, and gray bins, respectively. The RMSE shows the deviation between the simulated and empirical exceedance probabilities, and is hence dimensionless.

ceedance probabilities to each of the individual parameters in the VG (top panel) and Kosugi (bottom panel) model. Color coding is used to differentiate among the different parameters of each model. We separately display the results for the 2- (left column), 3- (middle column), and 5- (right column) parameter formulations of both models. We follow *Vrugt et al.* [2002] and derive these partial sensitivities, $\partial G_t/\partial x_i$ analytically by differentiation of the VG and Kosugi WRF models to each of its parameters. For example, for the 2-parameter formulation

of Kosugi, $\mathbf{x} = \{a_{\mathrm{K}}, b_{\mathrm{K}}\}$, the partial parameter sensitivities are given by

$$\begin{aligned} \partial G_t/\partial a_{\mathrm{K}} &= \tfrac{1}{\sqrt{2\pi}} a_{\mathrm{K}}^{-1} b_{\mathrm{K}}^{-1} \exp\left(\tfrac{\log(y_t/a_{\mathrm{K}})^2}{2b_{\mathrm{K}}^2}\right)^{-1} \\ \partial G_t/\partial b_{\mathrm{K}} &= \tfrac{1}{\sqrt{2\pi}} \log(y_t/a_{\mathrm{K}}) b_{\mathrm{K}}^{-2} \exp\left(\tfrac{\log(y_t/a_{\mathrm{K}})^2}{2b_{\mathrm{K}}^2}\right)^{-1}, \end{aligned} \tag{6.13}$$

and their numerical values are plotted in Figure 6.4B with a blue and red line, respectively using representative values of a_{K}, b_{K}, and $\mathbf{Y} = \{y_1, \ldots, y_N\}$. The results presented in Figure 6.4 highlight several important findings.

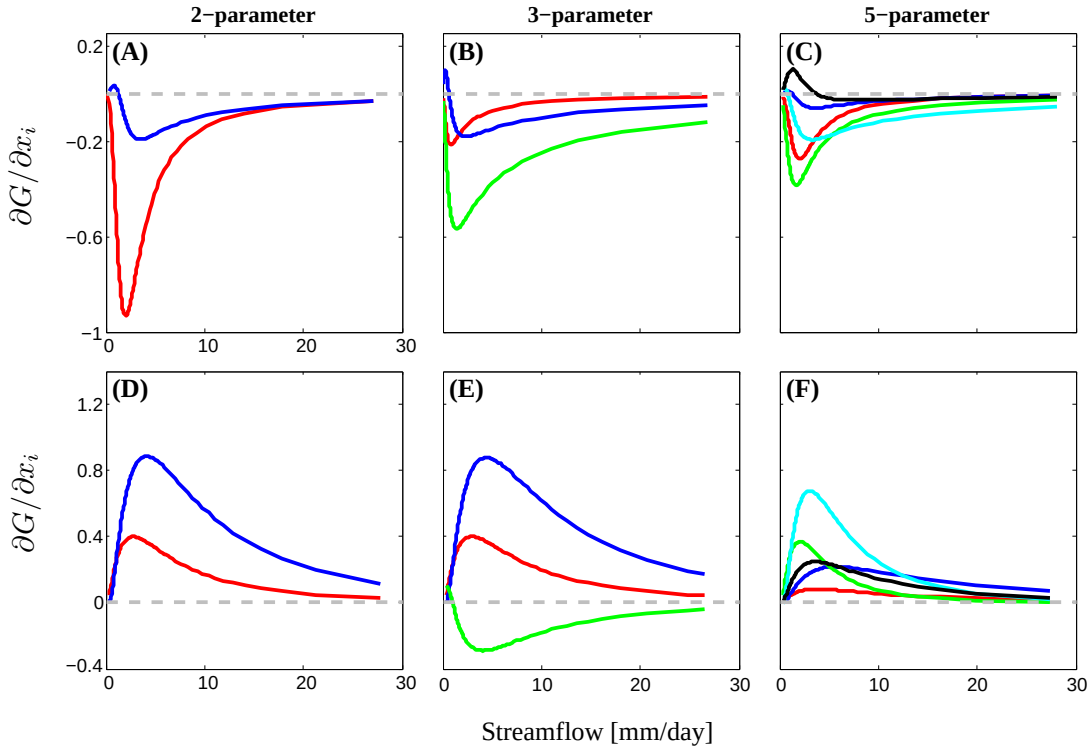


Figure 6.4: Partial parameter sensitivities of the VG (top panel) and Kosugi (bottom panel) models of the FDC (see Table 6.1) for a range of different streamflow values. Color coding is used to discern between the parameters a_{X} and $a_{1,\mathrm{X}}$ (red), b_{X} and $b_{1,\mathrm{X}}$ (blue), c_{X} and $a_{2,\mathrm{X}}$ (green), $b_{2,\mathrm{X}}$ (cyan), and $w_{1,\mathrm{X}}$ (black) of the different model formulations. The subscript X stands for VG and K. To help with visual interpretation, the gray dashed line depicts, for all flow levels, a zero sensitivity.

In the first place, the parameters of the VG and Kosugi FDC models appear most sensitive at lower streamflow values (0 - 10 mm/day), and are almost insensitive to high flow levels.

In other words, the values of the FDC parameters are determined primarily from low and intermediate flow levels (nondriven part of the hydrograph). Secondly, parameter sensitivity decreases with increasing model complexity. This finding is most apparent for the VG model and confirms the well known trade-off that exists between quality of fit and parameter identifiability. Indeed, the more parameters that are used to fit the empirical FDCs the more difficult it becomes to constrain their values. We will revisit this topic later using Bayesian inference with DREAM. Thirdly, the parameters reach their maximum sensitivity at approximately similar streamflow values. This finding is rather unfortunate as it makes both models vulnerable to parameter correlation (interaction). This is not particularly desirable if the fitting coefficients are used as summary statistics for diagnostic model evaluation . Finally, the partial parameter sensitivities appear jointly positive or negative (exceptions are the 3-parameter Kosugi and 5-parameter VG model) and one would therefore expect most (or all) of the parameters in each model to be correlated negatively. This can be easily verified if we consider a simple linear function with unknown slope and intercept. If we assume these two fitting coefficients to be larger than zero, then their partial sensitivities are both positive. This must induce a negative parameter correlation as an overestimation of the slope can be compensated for (somewhat) by a lower value of the intercept. A similar negative correlation is found where the slope and intercept are strictly negative. Thus if parameter uncertainty is of main concern then the 2-parameter formulations of VG and Kosugi are preferred as their fitting coefficients are most sensitive and thus best defined by fitting against the empirical FDCs.

We now provide a more detailed visual comparison of the observed and fitted FDCs. We select three watersheds with widely different rainfall-runoff response from the MOPEX data set and present their empirical FDCs in Figures 6.5 (Green river), 6.6 (Kankakee river), and 6.7 (Little river). The fit of the VG, Kosugi, Gumbel, Quimpo, GEV and Franchini and Suppo models is color coded and indicated with the black, blue, cyan, dark green, light green and gray lines, respectively. The top panel in each figure uses a linear scale of the observed

streamflow values, whereas the bottom panel displays the results using a logarithmic scale of the flow levels. The use of these two scales helps to illuminate differences among the models. The three different columns portray the results of the 2- (left: A,D), 3- (middle: B,E), and 5- (right: C,F) parameter formulations.

The FDCs of the three different watersheds plotted in Figures 6.5, 6.6 and 6.7 differ quite substantially from each other. Not only in terms of the range of their streamflow observations (y-axis) but also their functional shapes. The FDCs of the Green and Little rivers appear more similar and follow closely a strongly nonlinear, banana-shaped function when plotted using a linear streamflow scale. The Kankakee watershed on the contrary has an obvious right tail, which suggests the presence of an air-entry value in WRFs. Note that all these three different FDCs become "S"-shaped when a logarithmic scale is used for the streamflow values.

Visual comparison of each of the models against the data demonstrates that the proposed WRF formulations of the FDC are in much better agreement with their measured, empirical, counterparts than the existing FDC models used in the literature. This is most obvious in the bottom panel of each Figure when a logarithmic scale is used for the streamflow values. If we focus attention to the the 2-parameter FDC models, then the Kosugi model exhibits the best overall fit. This conclusion is also supported by our findings in Table 6.2, which shows that for about 60% of the MOPEX watersheds the Kosugi model is superior and achieves the lowest RMSE. Although the 2-parameter VG model nicely fits the FDCs of a very large number of watersheds, this model systematically overestimates the exceedance probabilities in the left tail of the FDC curve at the higher flow levels. This deficiency is particularly noticeable for the Little river (Figure 6.7) and highlights a structural weakness of the VG model for watersheds with a very large number of low and high flow events and relatively few streamflow observations in between.

The flow of the Little river falls rapidly from approximately 35 to 7 mm/day, then decreases

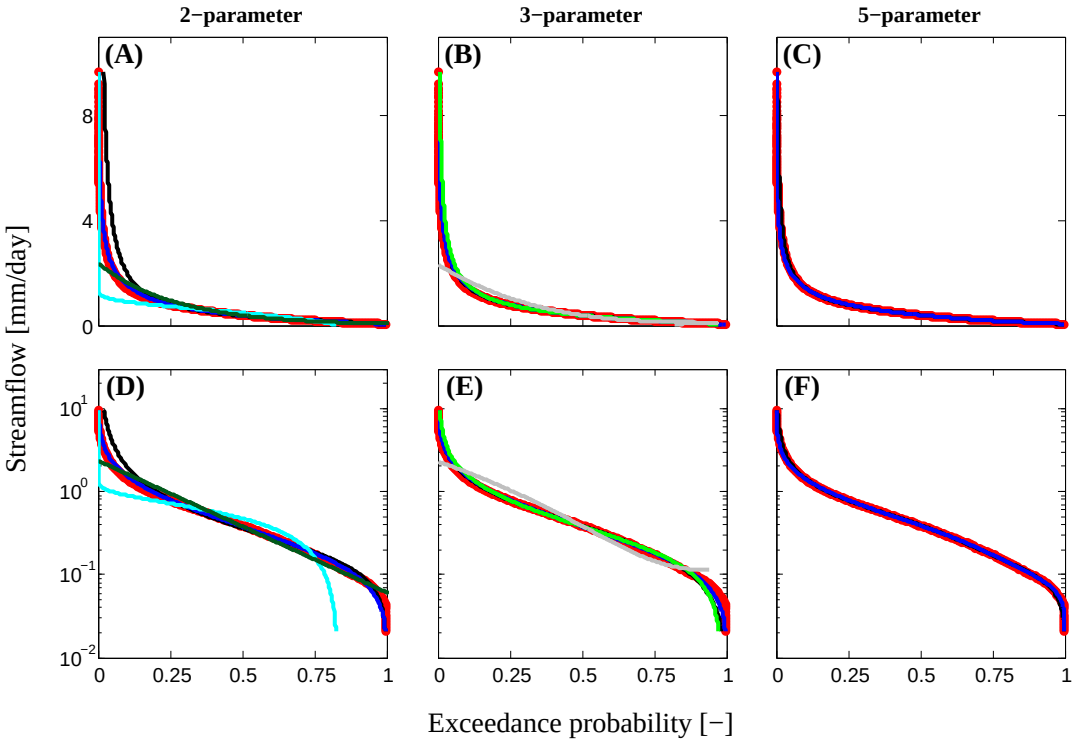


Figure 6.5: Comparison of the observed (red dots) and fitted (different lines) FDCs of the Green river watershed near Genesco, IL, USA. The top (A-C) panel uses a linear scale of the streamflow values, whereas the bottom (D-F) panel uses a logarithmic scale. The different columns represent different model complexities. Color coding is used to differentiate between the FDC models of Table 6.1: VG (black), Kosugi (blue), Gumbel (cyan), Quimpo (dark green), GEV (light green) and Franchini and Suppo (gray).

linearly to about 2 mm/day, and then smoothly reduces to values of zero. Although the 2-parameter formulations of VG and Kosugi cannot describe perfectly the FDC of this watershed, their 5-parameter variants (Figure 6.7C and F) are superior and fit adequately the exceedance probabilities for the entire range of streamflow values. This finding is important as it points to the presence of two or more alternating flow regimes of the Little river watershed. Indeed, this particular watershed alternates rapidly between a relatively dry and wet state, which, in part, is caused by the temporal characteristics of the rainfall (long dry periods interrupted by short but heavy precipitation events). During the long dry periods the runoff of this basin is made up entirely of baseflow originating directly from the groundwater reservoir. This flow regime changes dramatically during a storm event when a large fraction of water is transported directly to the stream network by surface runoff. This rapid change

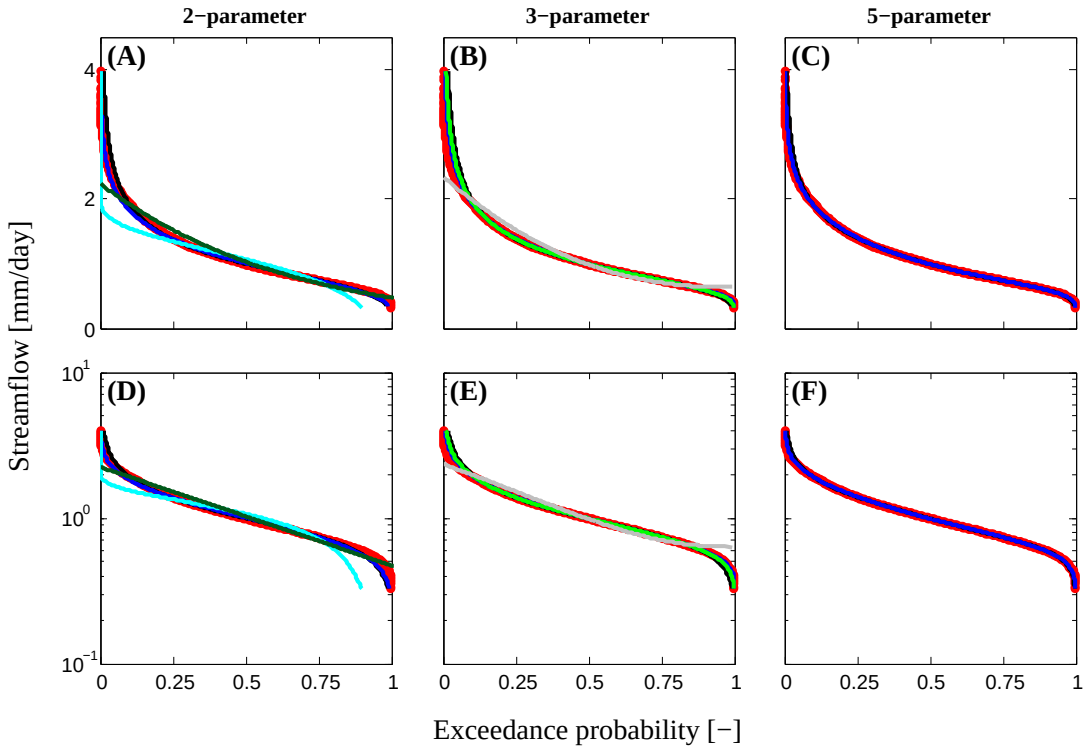


Figure 6.6: Comparison of the observed (red dots) and fitted (different lines) FDCs of the Kankakee river at Davis, IN, USA. The top (A-C) panel uses a linear scale of the streamflow values, whereas the bottom (D-F) panel uses a logarithmic scale. The different columns display the results of the 2 (left), 3 (middle) and 5 (right) parameter formulations.Color coding is used to differentiate between the FDC models of Table 6.1: VG (black), Kosugi (blue), Gumbel (cyan), Quimpo (dark green), GEV (light green) and Franchini and Suppo (gray).

in flow path can only be simulated adequately with a mixture of two (or more) WRFs. We conclude that if the 5-parameter formulations of the VG and Kosugi models much better fit the FDCs than their 2-, and 3-parameter formulations, the watershed likely experiences two or more alternating flow regimes.

The 3-parameter formulations of VG and Kosugi achieve a very similar fit to the data and their results can be considered nearly equivalent for these three watersheds. However, some deviations of the 3-parameter VG model are apparent at low flows (high exceedance probabilities), particularly for the Green and Little river watersheds. From the fitting results of these three watersheds we conclude that the Kosugi class of functions is best supported by the available FDC data.

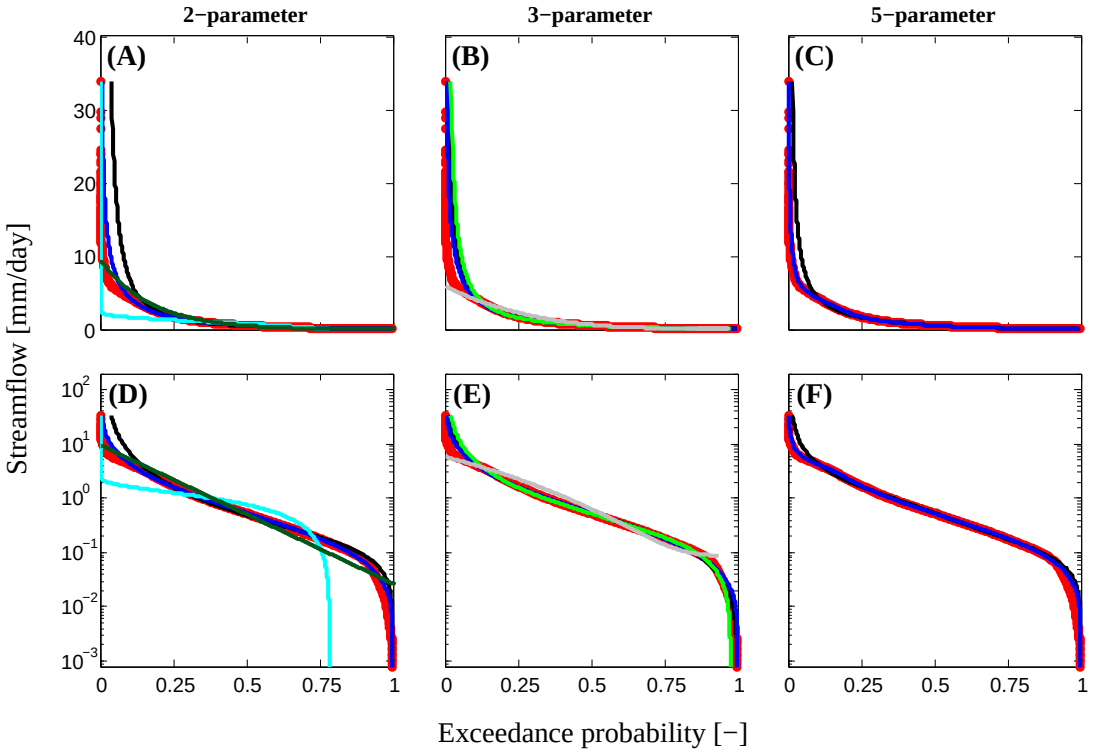


Figure 6.7: Comparison of the observed (red dots) and fitted (different lines) FDCs of the Little river near Horatio, AR, USA. A linear scale of the streamflow values is used in the top panel, whereas the bottom panel uses a logarithmic y-scale. The different columns display the results of the 2 (left), 3 (middle) and 5 (right) parameter formulations. The VG model is portrayed in black, and the Kosugi model in blue. The other FDC models proposed in the literature are displayed in cyan (Gumbel), dark green (Quimpo), light green (GEV) and gray (Franchini and Suppo).

We now proceed our analysis with an in-depth study of the interaction between the parameters of the WRF-based FDC formulations. Table 6.3 lists average linear correlation coefficients of the fitting coefficients of the VG and Kosugi models derived from MCMC simulation with DREAM using the 408 different watersheds of the MOPEX data set. These values are calculated by computing the mean of the absolute values of the linear correlation coefficients among the posterior parameter samples of each individual watershed. Absolute values have to be used for each watershed to avoid cancellation of positive and negative correlations. Of course, strong parameter correlation is not of much importance if quality of fit is the main concern in application of the VG and Kosugi FDC models, but becomes particularly important if the fitting coefficients of each model are used as summary metrics

for diagnostic analysis. Independent model parameters are also desired for regionalization studies when relating model parameters to one or more basin characteristics.

Table 6.3: Average linear correlation coefficients of the DREAM-derived posterior parameter samples for the 408 different watersheds of the MOPEX data set. The variable "X" that is used in subscript stands for K or VG respectively.

	2-parameter formulations									
	a_{VG}	b_{VG}				a_{K}	b_{K}			
a_{X}	1.00	0.81				1.00	0.16			
b_{X}	0.81	1.00				0.16	1.00			
	3-parameter formulations									
	a_{VG}	b_{VG}	c_{VG}			a_{K}	b_{K}	c_{K}		
a_{X}	1.00	0.89	0.96			1.00	0.44	0.46		
b_{X}	0.89	1.00	0.93			0.44	1.00	0.67		
c_{X}	0.96	0.93	1.00			0.46	0.67	1.00		
	5-parameter formulations									
	$a_{1,\mathrm{VG}}$	$b_{1,\mathrm{VG}}$	$a_{2,\mathrm{VG}}$	$b_{2,\mathrm{VG}}$	$w_{1,\mathrm{VG}}$	$a_{1,\mathrm{K}}$	$b_{1,\mathrm{K}}$	$a_{2,\mathrm{K}}$	$b_{2,\mathrm{K}}$	$w_{1,\mathrm{K}}$
$a_{1,\mathrm{X}}$	1.00	0.73	0.73	0.64	0.79	1.00	0.73	0.72	0.69	0.76
$b_{1,\mathrm{X}}$	0.73	1.00	0.65	0.62	0.77	0.73	1.00	0.69	0.62	0.70
$a_{2,\mathrm{X}}$	0.73	0.65	1.00	0.71	0.79	0.72	0.69	1.00	0.71	0.74
$b_{2,\mathrm{X}}$	0.64	0.62	0.71	1.00	0.75	0.69	0.62	0.71	1.00	0.69
$w_{1,\mathrm{X}}$	0.79	0.77	0.79	0.75	1.00	0.76	0.70	0.74	0.69	1.00

Results demonstrate that the 2-parameter formulation of Kosugi is preferred if parameter correlation is of paramount importance. The two parameters of this FDC model exhibit an average linear correlation coefficient of about 0.16. This correlation is negligibly small and of little concern for regionalization studies and diagnostic analysis. Quite interestingly, the two parameters of the VG model interact much stronger with correlation coefficient that exceeds the value of 0.8. Apparently, the 2-parameter formulation of Kosugi not only exhibits the highest qualify of fit of all models with two fitting coefficients (e.g. Table 6.2) but also the lowest parameter interaction. This model is preferred for diagnostic analysis and regionalization studies (discussed later).

The correlation among the fitting coefficients of the VG and Kosugi models increases significantly if the number of FDC parameters is enhanced. Indeed, for the 3- and 5-parameter variants of VG and Kosugi the correlation coefficients have increased to values of about 0.7 - 0.9 and 0.4 - 0.7, respectively. These formulations do fit better the observed FDCs (see Table 6.2) but at the expense of a significant increase in parameter correlation. Quite interestingly, the correlation coefficients of the 5-parameter VG model are lower than its 3-parameter counterpart. It is not immediately obvious what causes this reduction in parameter correlation. We leave this for future studies.

Now we have determined that the Kosugi model is preferred for FDC fitting and diagnostic analysis, we now present in Figures 6.8 and 6.9 the results of our 2- and 5-parameter formulations of this model for 12 different watersheds in the USA with quite different hydrologic responses to rainfall. This includes (A) SB Potomac river, WV, (B) Tygart Valley river, WV, (C) NF Holston river, VA, (D) Little river, AR, (E) Green river, IL, (F) Kankakee river, IN, (G) EF San Gabriel river, CA, (H) White river, WA, (I) Pemigewasset river, NH, (J) Little Pee Dee river, SC, (K) Licking river, KY, and (L) Genesee river, NY. The empirical FDCs exhibit quite different functional shapes but the fit of both Kosugi formulations can be considered excellent.

Some deviations of the 2-parameter Kosugi model remain apparent at the high flows (low exceedance probabilities), particularly for semi-arid and arid watersheds that are characterized by long dry periods and occasional but large precipitation events (convective thunderstorms). Fortunately, the 5-parameter Kosugi function fits adequately these watersheds as well - in large part because of its ability to separately describe with two different WRFs the behavior of the watershed at low and high flows. The easiest watersheds to fit are those that are relatively wet (lots of precipitation) and in which the groundwater reservoir contributes a large part of the total streamflow. The FDCs of such watersheds exhibit a relatively mild slope, and this is easy to describe with just two or three parameters. Examples includes the

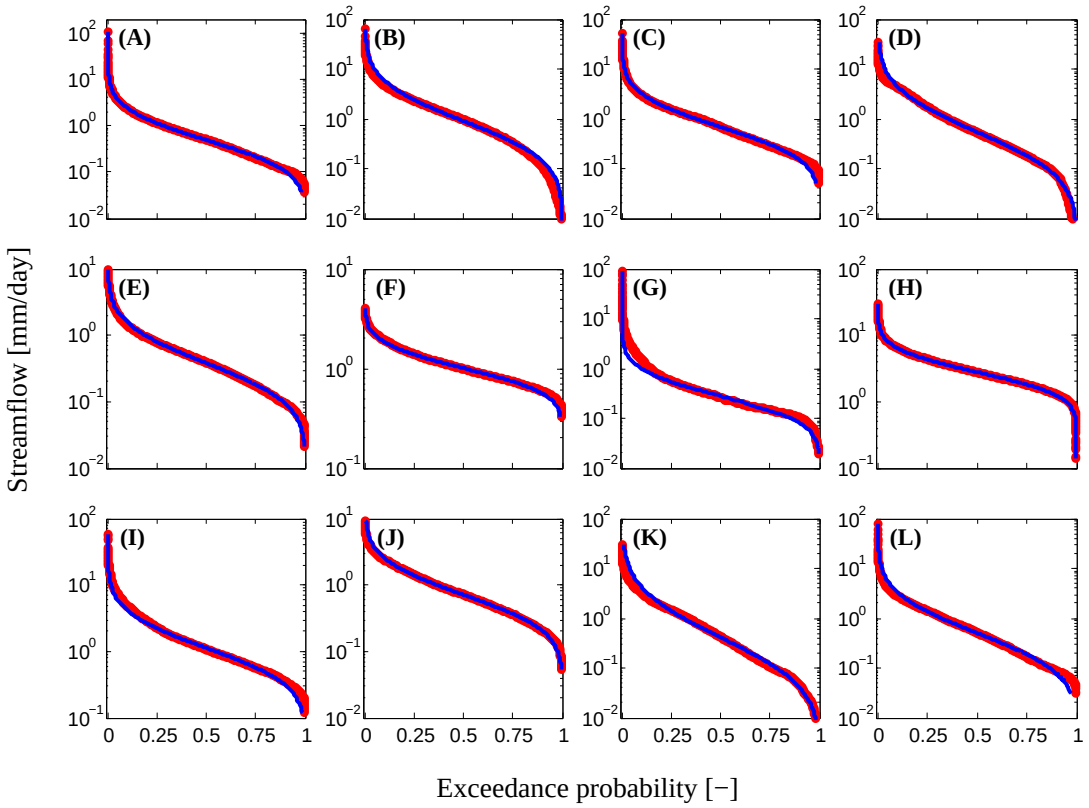


Figure 6.8: Comparison of the simulated FDCs of the 2-parameter Kosugi model (blue lines) against their observed counterparts (red dots) of the MOPEX data set. The different graphs correspond to the (A) SB Potomac, (B) Tygart Valley, (C) NF Holston, (D) Little, (E) Green, (F) Kankakee, (G) EF San Gabriel, (H) White, (I) Pemigewasset, (J) Little Pee Dee, (K) Licking, and (L) Genesee river basins, respectively, which are found in the states of West Virginia (A,B), Virginia, Arkansas, Illinois, Indiana, California, Washington, New Hampshire, South Carolina, Kentucky, and New York.

Kankakee (F), White (H), and Little Pee Dee (J) river basins.

We now focus our attention on model uncertainty, and on the 95% ranges of the 5-parameter Kosugi model for the (A) SB Potomac, (B) Tygart Valley, (C) NF Holston, (D) Little, (E) Green, (F) Kankakee, (G) EF San Gabriel, (H) White, (I) Pemigewasset, (J) Little Pee Dee, (K) Licking, and (L) Genesee river basins. We plot separately the effects of parameter (dark gray) and model (light gray) uncertainty in Figure 6.10. The fit to the data is excellent and this makes it difficult to visualize clearly the results. The parameter and total model uncertainty is rather small, and almost covered completely by the data of the empirical FDCs of each watershed. The exception to this are the NF Holston (C), Little (D), and Genesee

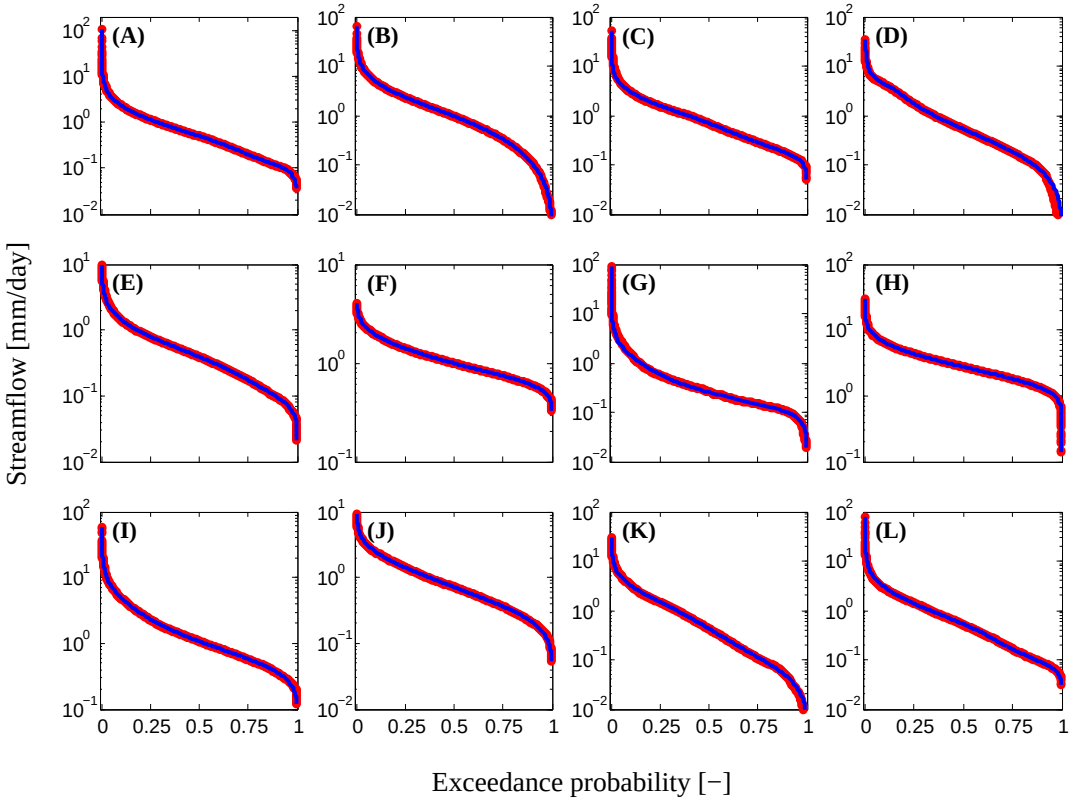


Figure 6.9: Comparison of the simulated FDCs of the 5-parameter Kosugi model (blue lines) against their empirical counterparts (red dots) of the MOPEX data set. The different graphs pertain to the (A) SB Potomac, (B) Tygart Valley, (C) NF Holston, (D) Little, (E) Green, (F) Kankakee, (G) EF San Gabriel, (H) White, (I) Pemigewasset, (J) Little Pee Dee, (K) Licking, and (L) Genesee river basins respectively which are located in the states of West Virginia (A,B), Virginia, Arkansas, Illinois, Indiana, California, Washington, New Hampshire, South Carolina, Kentucky, and New York.

(L) river basins. The simulation uncertainty ranges are somewhat larger as the empirical FDCs are not fitted perfectly close at all times. Indeed, the 95% parameter uncertainty is clearly visible in plots (C) and (L) and show the small deviation of the fitted function from the empirical FDC.

The 2-parameter lognormal WRF model of Kosugi has the advantage that its parameters can be directly related to the pore size distribution. In other words, the parameters should have a physical interpretation. We verify this and calculate the linear correlation coefficients between the five parameters of the Kosugi model and the MOPEX basin characteristics. Correlations are relatively low and range between values of $r = -0.5$ and $r = 0.5$. The

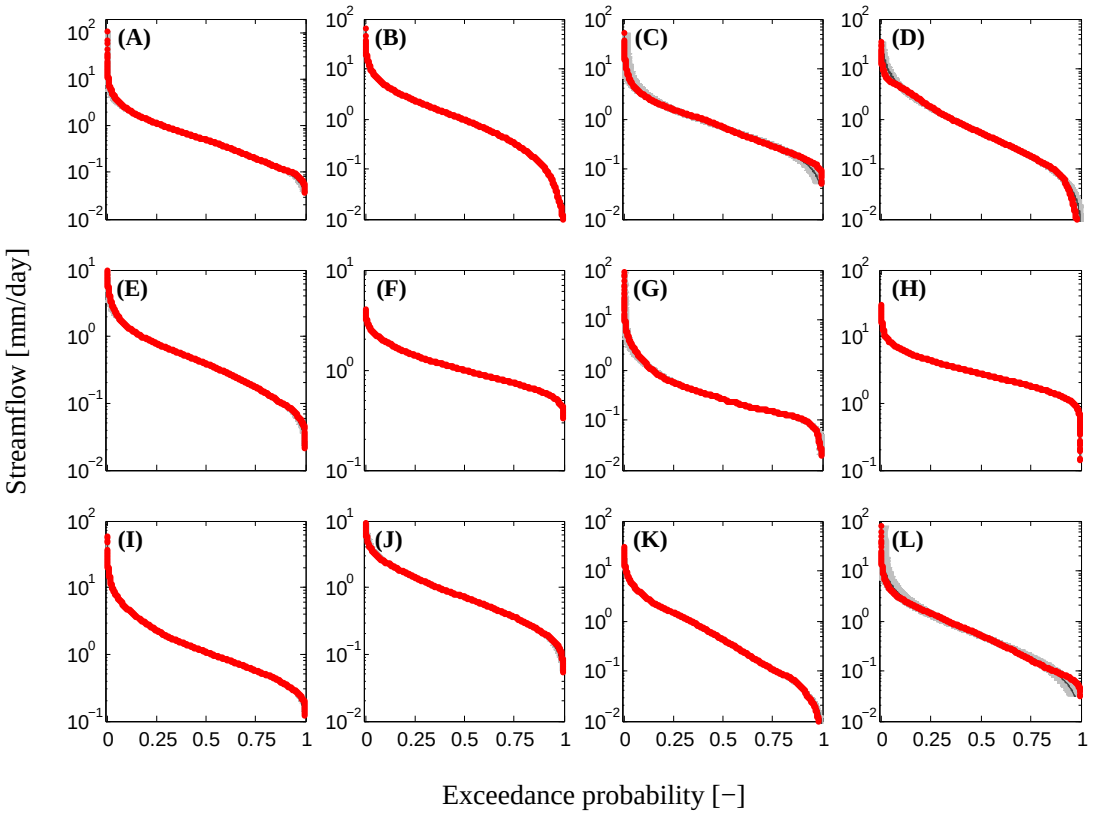


Figure 6.10: 95% simulation uncertainty ranges of the 5-parameter Kosugi model for the twelve different watersheds of Figure 6.8. The observed data are indicated with the red dots, whereas the 95% parameter and total uncertainty are displayed using the dark and light gray regions, respectively. The fit to the empirical FDCs is excellent and the parameter and model uncertainty therefore negligibly small.

two parameters of the original Kosugi function, however, show a much higher correlation with the physical and climatic characteristics of the watersheds. For instance, our analysis demonstrates that the parameters a_{K} and b_{K} are most related to climatic variables (mean annual precipitation and potential evapotranspiration) and reach linear correlation coefficients of approximately 0.8. The first parameter of the 2-parameter Kosugi model, a_{K}, shows a linear correlation coefficient $r = 0.9974$ to median streamflow y_{Mdn} for the 408 analysed watersheds (see Figure 6.11). This opens up new possibilities for regionalization and prediction in ungaged basins using for instance, remote sensing precipitation data products.

Now we have established an almost perfectly linear correlation between the median streamflow level and the value of parameter a_{K} of the 2-parameter Kosugi model, we can readily

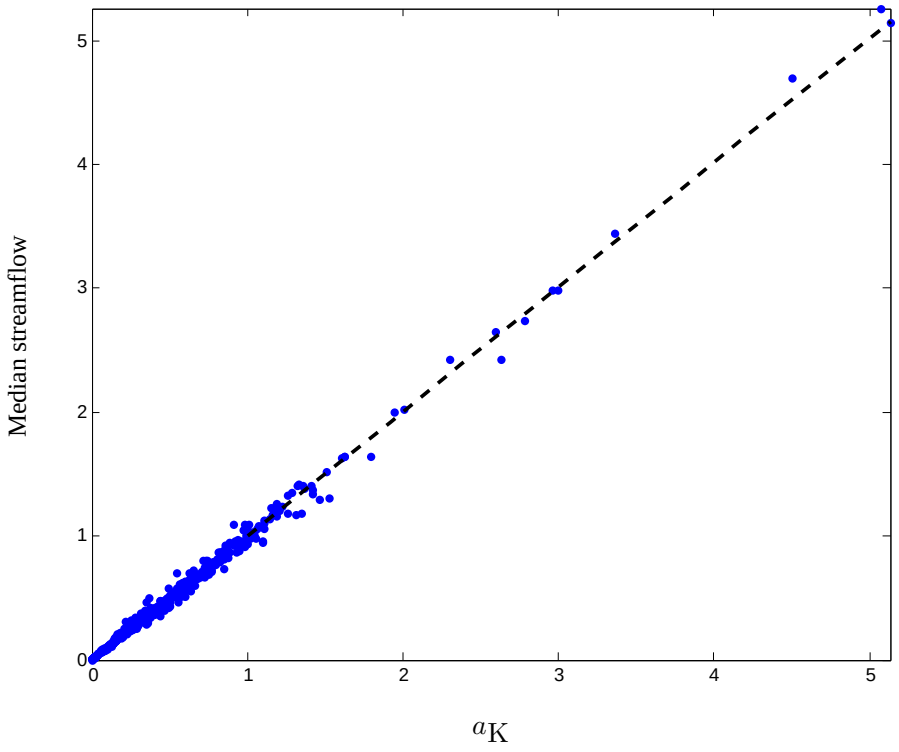


Figure 6.11: Scatter plot of first parameter of the original Kosugi FDC model, a_{K}, vs median streamflow, y_{Mdn}, for 408 watersheds of the MOPEX data set. a_{K} and y_{Mdn} are correlated with a linear correlation coefficient $r = 0.9974$ and a line $y_{\mathrm{Mdn},i} = 1.0061 a_{\mathrm{K},i} - 0.0061$ describes the relationship between them.

fit this model to the observed FDC using the following steps:

1. Set a_{K} equivalent to the median of the streamflow values, $a_{\mathrm{K}} = g(\tilde{\mathbf{Y}})$, where $g(\cdot)$ is the median operator, and $\tilde{\mathbf{Y}} = \{\tilde{y}_1, \ldots, \tilde{y}_N\}$ is a $N \times 1$ vector with observed streamflow.

2. Assume an initial value for b_{K}, say $b_{\mathrm{K}} = 1$ and then compute G (size $N \times 1$) from Equation 6.10 using $\tilde{\mathbf{Y}}$ and $c_{\mathrm{K}} = 0$.

3. Compute the partial sensitivity, $\partial G / \partial b_{\mathrm{K}}$ from Equation 6.13 using $\tilde{\mathbf{Y}}$, the optimized value for a_{K} from step 1, and $b_{\mathrm{K}} = 1$. This partial sensitivity is also referred to as the Jacobian, J and is of size $N \times 1$.

4. Now we can estimate b_{K} from a single step with Gauss Newton:

$$\hat{b}_{\mathrm{K}} = b_{\mathrm{K}} + \mathrm{inv}(J^T J) J^T (G - \hat{G}), \tag{6.14}$$

where G and $\hat{G}$ are the observed and Kosugi predicted exceedance probabilities, respectively, and $\hat{b}_{\mathrm{K}}$ signifies the optimized value of b_{K}.

We can now use the optimized values of a_{K} and b_{K} with Equation 6.10 to predict the exceedance probability, G_t for a given flow level, y_t (assuming $c_{\mathrm{K}} = 0$). Alternatively, we can invert Equation 6.10

$$y_t = a_{\mathrm{K}} \exp\big(\sqrt{2} b_{\mathrm{K}} \mathrm{ierfc}(2G_t)\big), \tag{6.15}$$

where ierfc denotes the inverse complementary error function, and predict the flow level for a given exceedance probability, G_t. Note, the correlation between a_{K} and the median streamflow level is not perfectly linear ($r = 0.997$). We can therefore get a somewhat improved fit to the empirical FDC if we fine tune a_{K}. This is easily done by using Equation 6.14 for a_{K} and b_{K} jointly.

We conclude the results section with some final remarks: (1) The fitting of the proposed FDC models is not without problems. We therefore recommend to use different starting points with a local search method to minimize chances of premature convergence to the wrong values of the FDC fitting coefficients, and (2) The proposed WRF-based functions demonstrate low sensitivity to high flows (Figure 6.4), and hence some care should be exercised if these models are used to estimate the exceedance probability of extreme events. One way to mitigate this problem is to define the objective function of Equation 6.12 in the streamflow space rather than exceedance probability space. This would require the functions of Table 6.1 to be inverted so that they predict the streamflow value, y_t for each observed exceedance

probability, G_t.

6.4 Conclusions

The FDC has been used to characterize information regarding catchment streamflow variability and flow regime in various hydrologic design and management studies. While several physically-based and probabilistic/mathematical functions have been used to model the empirical FDCs, they fail to characterize adequately the mid-section and tails of the FDC at low and high exceedance probabilities, respectively, and the large differences between the FDCs of watersheds with contrasting hydrologic behaviors. This chapter extends the work of *Vrugt and Sadegh* [2013] and introduces a new class of closed-form parametric expressions to closely characterize the FDCs of arid, semi-arid, and wet catchments. These continuous functions, equivalent to the WRFs of VG and Kosugi, are derived from the field of vadose zone hydrology and have been used widely to describe the water characteristic of variably saturated soils. Three different formulations of the VG and Kosugi models have been introduced to mimic empirical FDCs. These functions differ in their structural complexity and include between two to five coefficients whose values need to be derived by fitting against the observed FDCs.

Application of the WRFs to 408 watersheds of the MOPEX data set demonstrates that the VG and Kosugi models closely match the empirical FDCs of arid, semi-arid, and wet catchments. The proposed parametric expressions exhibit a superior performance over commonly used FDC models in the hydrologic literature. If quality of fit is of paramount importance then the 5-parameter Kosugi model is preferred as it provides the closest fit to the empirical FDCs. What is more, this mixture formulation of two WRFs can help to discern different stable states of the catchment. If the fitting coefficients of the FDC model are to be used for regionalization or as summary statistics in diagnostic analysis then the 2-parameter for-

mulation of Kosugi is preferred. The parameters of this model exhibit a negligibly small correlation and are therefore suitable metrics that are well defined by calibration against the observed FDCs. We conclude that a maximum of four-five parameters is sufficient to describe the large range of FDCs of the MOPEX data set.

Chapter 7

Summary Statistics From Training Images as Prior Information in Probabilistic Inversion

This Chapter is currently in Press in Geophysical Journal International. Citation: Lochbühler, T., J.A. Vrugt, M. Sadegh, and N. Linde (2015), Summary statistics from training images as prior information in probabilistic inversion, *Geophysical Journal International*, In Press.

7.1 Introduction

Geophysical inversion seeks to infer representative estimates of spatially distributed subsurface properties, given direct or indirect measurements that are sensitive to those properties. For a specified model parameterization, each set of parameter values is referred to as a model. In probabilistic inversions, the probability of a certain model given the data is expressed as the product of a prior distribution and a likelihood function using Bayes Theorem. The prior distribution is a probability density function (pdf) that summarizes all information available about the parameters of interest before any data is collected, whereas the likelihood function quantifies the probability that the actual data with its measurement errors has been generated by a proposed model. The product of the prior distribution and likelihood function leads to a posterior distribution, which summarizes the statistical distribution of the model parameters. Unfortunately, in most practical applications we cannot derive this posterior distribution by analytical means. We therefore resort to Monte Carlo simulation to generate samples from the posterior distribution. The most popular of such sampling methods is the Random Walk Metropolis (RWM) algorithm (e.g., *Chen et al.* [2000]), a Markov chain Monte Carlo (MCMC) simulation method that has found widespread application and use in many different fields of study. Such MCMC methods have several advantages over deterministic, gradient-driven optimization methods in that they (1) do not require linearization of the inverse problem and thus can handle highly non-linear problems and (2) provide an ensemble of models drawn from the posterior distribution that can be used to characterize parameter uncertainty. Recent applications of MCMC in geophysics include inversions of electromagnetic [*Rosas-Carbajal et al.*, 2014], seismic [*Bodin and Sambridge*, 2009, *Hong and Sen*, 2009], and ground-penetrating radar data [*Scholer et al.*, 2012], as well as multi-physics inversions [*Bosch et al.*, 2006, *Irving and Singha*, 2010]. *Sambridge and Mosegaard* [2002] present a comprehensive but somewhat dated review on Monte Carlo methods in geophysical inverse modeling.

A general problem of parameter estimation in geophysics is that there exists an infinite number of models that can explain the data within their error bounds *Backus and Gilbert* [1970], but most of them are unrealistic or even unphysical. To overcome this ambiguity and reduce the number of possible models, it is common practice to impose data independent constraints on the model structure. Constraints can be imposed by choosing model parameterizations where the model space is bounded to exclude undesired models (e.g., *Lochbühler et al.* [2014a]), or by favoring models based on their agreement with a reference model, the model roughness [*Constable et al.*, 1987, *Rosas-Carbajal et al.*, 2014], the fit to a multi-Gaussian geostatistical model [*Maurer et al.*, 1998, *Johnson et al.*, 2007] or the structural similarity to another model in joint inversions (e.g., *Gallardo and Meju* [2004], *Linde et al.* [2006]). These constraints are accounted for by including penalty terms in the objective function (deterministic inversion) or by using an explicit prior pdf (probabilistic inversion). Alternatively, geostatistical inversion approaches (e.g., *Irving and Singha* [2010], *Mariethoz et al.* [2010a], *Hansen et al.* [2012c]) adopt a constrained prior in the form of a variogram and estimate the parameter values by sequential resimulation of parts of the model domain so that the underlying geostatistics are honored.

The geostatistical inversion approach has been extended recently to consider priors that account for higher order spatial statistics, so-called multiple-point statistics (MPS). The concept of MPS is becoming increasingly popular in the geostatistics and geophysics communities, since it allows for models that are constrained by detailed conceptual information about the expected geology at the site of interest [*Caers and Zhang*, 2004, *González et al.*, 2007, *Huysmans and Dassargues*, 2009, *Lochbühler et al.*, 2014b]. The basis for MPS is a training image (TI), which is a graphical representation of a conceptual geological model that features the lithologies and structural patterns that could be expected at a given site (e.g., *Strebelle* [2002]). The MPS formulation of the spatial variability of parameter values allows for description of features that covariance-based, two-point statistics are unable to describe accurately such as curvilinear and discrete continuous structural elements. A TI thus offers

strong prior information. It can be used in geostatistical inversions where MPS simulations are conditioned to geophysical data by sequential updating of groups or blocks of cells [*Mariethoz et al.*, 2010a, *Hansen et al.*, 2012c, *Cordua et al.*, 2012, *Hansen et al.*, 2012a,b]. The information content of the TI drastically reduces the degrees of freedom compared with the number of cells in the model domain, and the resulting posterior models are visually very similar to the TI (e.g., *Cordua et al.* [2012]). In these approaches, the sampling efficiency strongly depends on the number of updated cells and it can be difficult to fit large data sets with realistic observational errors (e.g., *Irving and Singha* [2010], *Cordua et al.* [2012]).

We here present an inversion strategy where TI realizations, that is, images comprising the multiple-point statistics represented by the TI, are used to define case-specific model parameterizations and prior density functions. Models are parameterized by coefficients of their truncated discrete cosine transform (DCT), with coefficients based on analysis of the TI realizations in the DCT domain [*Jafarpour et al.*, 2009]. The DCT coefficients that are necessary to describe structures that frequently occur in the TI will thus be part of the model parameterization and such structures can be reproduced in the posterior models. Apart from the model parameterization, we also extract from the TI realizations the distributions of global measures of model morphology, what we refer to as summary statistics. These distributions are translated into prior distributions to evaluate the prior probability of all proposed models.

Our inversion strategy is conceptually different from the geostatistical inversion approach described above, as all model parameters are sampled independently. There is also a clear distinction with respect to regularized inversions (e.g., *Constable et al.* [1987], *Rosas-Carbajal et al.* [2014], *Maurer et al.* [1998]), in that information on the model structure is not imposed by enforcing intercellular correlation of a certain form, but instead by defining a prior on global, location-independent summary metrics. If, for example, the extracted summary metric is the global roughness, the prior probability of a proposal model depends on the

global roughness distribution of the TI realizations. This distribution is unlikely to be centered around zero as imposed in classical smoothness-constrained inversions.

In the remainder of this chapter, we present the individual building blocks of our inversion strategy (Section 7.2) and illustrate our preliminary findings for two different case studies with different geological settings (Section 7.3) using travel time observations from synthetic crosshole ground-penetrating radar (GPR) experiments. Section 7.4 discusses the advantages and limitations of the proposed methodology. Finally, in Section 7.5 we provide our conclusions.

7.2 Methodology

7.2.1 Sparse model parameterization

In applied geophysics, the model space is typically parameterized by pixels (or voxels) in 2-D (3-D) Cartesian grids. For large model domains or fine grid discretizations, this quickly results in several thousands to millions of model parameters. Since the efficiency of MCMC algorithms decreases sharply with increasing number of independently sampled parameters, high resolution Cartesian parameterization is a daunting and CPU-intensive task, particularly in the absence of a prescribed spatial correlation structure. Fortunately, the use of sparse representations in the frequency-amplitude domain can help to significantly reduce the dimensionality of the model space while maintaining a large degree of fine-scale information. Alternative parameterizations of the model space include the discrete cosine [*Jafarpour et al.*, 2009, 2010, *Linde and Vrugt*, 2013, *Lochbühler et al.*, 2014a] and the wavelet transform [*Jafarpour*, 2011, *Davis and Li*, 2011]. Here we use the discrete cosine transform (DCT, *Ahmed et al.* [1974]), more precisely, DCT-II in 2-D, which for a uniformly discretized model

$\mathbf{A} \in \mathcal{R}^{N_x \times N_z}$ is given by

$$B(k_x, k_z) = \alpha_{k_x} \alpha_{k_z} \sum_{x=0}^{N_x-1} \sum_{z=0}^{N_z-1} A(x, z) \cos \frac{\pi(2x+1)k_x}{2N_x} \cos \frac{\pi(2z+1)k_z}{2N_z}, \tag{7.1}$$

where

$$\alpha_{k_x} = \begin{cases} \frac{1}{\sqrt{N_x}}, & k_x = 0 \\ \\ \sqrt{\frac{2}{N_x}}, & 1 \leq k_x \leq N_x - 1, \end{cases} \tag{7.2}$$

and

$$\alpha_{k_z} = \begin{cases} \frac{1}{\sqrt{N_z}}, & k_z = 0 \\ \\ \sqrt{\frac{2}{N_z}}, & 1 \leq k_z \leq N_z - 1, \end{cases} \tag{7.3}$$

The matrix $\mathbf{B}$ contains the DCT coefficients. In the absence of prior information about the information content of each individual DCT coefficient, model compression can be adopted by truncating the high-frequency terms, and thus focusing on the low-frequency coefficients only [*Jafarpour et al.*, 2009, *Linde and Vrugt*, 2013]. In this study, we use prior information from TIs to determine case-specific sparse model parameterizations, which we hereafter conveniently refer to as TI-based parameterizations. Following *Jafarpour et al.* [2009], the dominant transform coefficients (i.e., those coefficients that contribute the most to represent the expected subsurface structures) are determined by the following steps:

1. create a large set of realizations of a known training image using a multiple-point statistics simulator (so-called TI realizations);

2. calculate the 2-D discrete cosine transform of all TI realizations;

3. compute the arithmetic mean of the absolute value of each transform coefficient;

4. select the n coefficients for which the mean value is largest, where n is a default value set by the user;

5. store the coordinates in the DCT space of the n dominant coefficients.

Jafarpour et al. [2009] used the reciprocals of the n dominant DCT coefficients to build a weighting matrix $\mathbf{W}$ that was used as a regularization term in a deterministic inversion. In the present study, the dominant transform coefficients are the model parameters that we seek to estimate, while all other coefficients are set to zero. The DCT was preferred above the wavelet transform, as for the investigated cases the DCT had superior compression power to achieve comparable model quality.

The choice of the number of considered DCT coefficients, n, is a trade-off between the desired resolution and the available computational budget. The quality of the compression can be quantified by the peak signal-to-noise ratio (PSNR)

$$\mathrm{PSNR} = 10 \log_{10} \left(\frac{\max(\mathbf{A})^2}{\frac{1}{N_x N_z} \sum_{x=1}^{N_x} \sum_{z=1}^{N_z} \left[A(x,z) - A'(x,z) \right]^2} \right), \tag{7.4}$$

where $\mathbf{A}$ and $\mathbf{A}'$ are the uncompressed and compressed images, respectively. The PSNR is a common measure in image processing (e.g., *Huynh-Thu and Ghanbari* [2008]). For the two cases presented herein, we use $n = 100$ and $n = 150$, respectively, resulting in PSNR values between 28 and 39 dB. This is well above common PSNR thresholds in image processing (e.g., *Welstead* [1999]).

7.2.2 Markov chain Monte Carlo sampling with DREAM(ZS)

We seek to infer the posterior distribution of a set of model parameters $\mathbf{m}$ from observed data $\mathbf{d}$. In a probabilistic sense, the relation between the model and the data can be expressed

by Bayes Law

$$p(\mathbf{m}|\mathbf{d}) = \frac{p(\mathbf{m})p(\mathbf{d}|\mathbf{m})}{p(\mathbf{d})} = \frac{p(\mathbf{m})L(\mathbf{m}|\mathbf{d})}{p(\mathbf{d})}, \tag{7.5}$$

where $p(\mathbf{d}|\mathbf{m})$, the posterior pdf of the model given the data, is our main subject of interest. The other entities in Equation 7.5 are $p(\mathbf{m})$, the prior distribution of the model parameters, $L(\mathbf{m}|\mathbf{d}) \equiv p(\mathbf{d}|\mathbf{m})$, the likelihood of the parameters given the data and $p(\mathbf{d})$, the evidence. The latter term is generally difficult to estimate in practice, but if the model parameterization is fixed, all relevant statistical information (mean, standard deviation, etc.) can be extracted from the unnormalized distribution

$$p(\mathbf{m}|\mathbf{d}) \propto p(\mathbf{m})L(\mathbf{m}|\mathbf{d}). \tag{7.6}$$

In this study, the data $\mathbf{d}$ constitute the radar travel times from crosshole GPR experiments. The model vector $\mathbf{m}$ includes the dominant DCT coefficients for a given log-porosity field $\boldsymbol{\varphi}$, $\mathbf{m} = \tau_{\text{DCT}}\log(\boldsymbol{\varphi})$ and a set of petrophysical parameters that are necessary to solve the forward problem.

The prior $p(\mathbf{m})$ describes the distribution of the model parameters before any data $\mathbf{d}$ have been collected and assimilated (e.g., *Chen et al.* [2000]). The formulation of the prior distribution based on summary statistics inferred from TI realizations is a central theme of this study and will be described in detail later (Section 7.2.3).

The likelihood function $L(\mathbf{m}|\mathbf{d})$ measures the distance between the measured data, $\mathbf{d}$, and the corresponding values predicted by the proposed model, $\mathbf{d}^{\text{pred}}$. If we assume the error residuals to be uncorrelated and Gaussian distributed, the likelihood function, $L(\mathbf{m}|\mathbf{d})$, is given by

$$L(\mathbf{m}|\mathbf{d}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2}\frac{(d_i^{\text{pred}}(\mathbf{m}) - d_i)^2}{\sigma_i^2}\right], \tag{7.7}$$

where N denotes the number of data points and σ_i represents the measurement error standard deviation of the i-th data point.

We apply a hierarchical Bayes scheme (e.g., *Malinverno and Briggs* [2004]) and estimate a global relative error level $\sigma_{\text{rel}} = \sigma_i/d_i$, jointly with the parameters $\mathbf{m}$. By doing so, we also account for epistemic (model structural) and model parameterization errors, as long as these are independent and well described by the Gaussian likelihood function in Equation 7.7 (e.g., *Schoups et al.* [2010b], *Rosas-Carbajal et al.* [2014]). Relative errors are used herein, but the perhaps more common assumption of absolute errors would also be straightforward to implement and lead to similar results. The posterior pdf is sampled with the DREAM$_{\text{(ZS)}}$ algorithm [*ter Braak and Vrugt*, 2008, *Vrugt et al.*, 2009a, *Laloy and Vrugt*, 2012]. We here give a brief description of the sampling scheme and refer to *Laloy and Vrugt* [2012] for a detailed description of the algorithm. In short, DREAM$_{\text{(ZS)}}$ is an adaptive MCMC algorithm that runs $K(K > 2)$ different chains in parallel and creates jumps in each chain using a fixed multiple of the difference of two states sampled from an archive of past model states. This archive is growing progressively as the sampling evolves and diminishing adaptation ensures convergence to the appropriate target distribution. We refer to *ter Braak and Vrugt* [2008] for an in-depth discussion concerning ergodicity and convergence properties of this algorithm. If the position of the i-th chain is given by $\mathbf{m}^i$, then new proposal models, $\mathbf{m}^i_{\text{prop}}$ are calculated using

$$\mathbf{m}^i_{\text{prop}} = \mathbf{m}^i + \mathbf{\Delta}^i. \tag{7.8}$$

The DREAM$_{\text{(ZS)}}$ algorithm implements subspace sampling in which only a subset of all dimensions is periodically updated. The dimensions to be updated (indexed j) are determined for each step based on a series of crossover values (details given by *Vrugt et al.* [2009a]). The number of updated dimensions, d' , lies within 1 and the total number of dimensions, d. The

respective dimensions are updated by a proposal jump

$$\boldsymbol{\Delta}_j^i = (\mathbf{1}_{d'} + \boldsymbol{e}_{d'})\gamma_{d'}\left[\boldsymbol{z}_j^{r1} - \boldsymbol{z}_j^{r2}\right] + \boldsymbol{\epsilon}_{d'}, \tag{7.9}$$

where $\mathbf{1}_{d'}$ denotes a unit vector of length d', $\boldsymbol{z}^{r1}$ and $\boldsymbol{z}^{r2}$ are sampled from the external archive of model states, $\mathbf{Z}$, and $r1$ and $r2$ are randomly chosen members of $\mathbf{Z}$. The variables $\boldsymbol{e}_{d'}$ and $\boldsymbol{\epsilon}_{d'}$ add stochastic fluctuations to ensure ergodicity and are drawn from $U_{d'}(-b, b)$ and $N_{d'}(0, b^*)$, respectively, where b and b^* are small compared to the width of the target distribution. The jump rate, $\gamma_{d'} = 2.4/\sqrt{2d'}$, depends on the number of updated dimensions d' [*Vrugt et al.*, 2009a] and controls the dissimilarity between subsequent states. Each fifth proposal, the jump rate is temporarily set to 1 to facilitate direct jumps between disconnected posterior modes.

The acceptance probability of the proposed model is determined using the Metropolis ratio (e.g., *Mosegaard and Tarantola* [1995])

$$\alpha = \min\left\{1, \frac{p(\mathbf{m}_{\text{prop}}^i|\mathbf{d})}{p(\mathbf{m}^i|\mathbf{d})}\right\}. \tag{7.10}$$

If the posterior probability of the proposed model, $p(\mathbf{m}_{\text{prop}}^i|\mathbf{d})$, is larger than that of the current position of the chain, $p(\mathbf{m}^i|\mathbf{d})$, the probability to move to this new state is one. Otherwise the acceptance probability is given by the ratio of $p(\mathbf{m}_{\text{prop}}^i|\mathbf{d})$ and $p(\mathbf{m}^i|\mathbf{d})$. For the sake of numerical stability, the likelihood function and the prior probability are evaluated in log-space (denoted $l(\mathbf{m}|\mathbf{d})$ and $\omega(\mathbf{m})$, respectively), and the acceptance probability to move from one state to the next becomes

$$\alpha = \min\left\{1, \exp\left(l(\mathbf{m}_{\text{prop}}^i|\mathbf{d}) + \omega(\mathbf{m}_{\text{prop}}^i) - l(\mathbf{m}^i|\mathbf{d}) - \omega(\mathbf{m}^i)\right)\right\}. \tag{7.11}$$

7.2.3 Summary statistics from training image realizations

Training images (TI) are 2- or 3-D images, digitally created or drawn by hand, that represent conceptual geological information of the area of interest such as characteristic structural elements and lithofacies [*Strebelle*, 2002, *Hu and Chugunova*, 2008]. Such images are typically used when the subsurface is described by repeating structures, for example, meandering and intersecting channels in the case of fluvial or turbiditic reservoirs. In geostatistical modeling, the purpose of a TI is to represent the spatial dependencies between a set of points larger than 2, that is, the multiple-point statistics (MPS).

Our method relies on a TI that captures the expected spatial patterns and lithofacies information at the study site, which we refer to as geological prior information. Each lithological unit of the TI is assigned a different porosity value using information from outcrop data or borehole cuttings. Then, an ensemble of different geostatistical realizations of the TI is generated using the simulation algorithm *DeeSse*, an improved and commercialized version of the original algorithm proposed by *Mariethoz et al.* [2010b]. *DeeSse* simulations are based on direct sampling of patterns from the TI to sequentially simulate all pixel values in the realization. Patterns are defined by a set of neighboring pixels that have already been assigned values in previous simulation steps. The pixel combination defined by the pixel values and lags of the neighboring pixels to the target pixel, that is, the pixel to be simulated, builds a pattern. The TI is scanned for such a pattern and once found, the pixel values of the TI are copied to the realization. This pixel-wise copy-paste scheme preserves the multiple-point statistics of the TI. The size of the neighborhood (number of adjoining pixels used) determines the order of spatial statistics of the TI that is honored in each realization. Indeed, if only one neighbor is used, this procedure is equivalent to regular variogram sampling. Simulation quality and computational cost of the *DeeSse* algorithm depend on a set of controlling algorithmic parameters, which we chose following the recommendations made by *Meerschman et al.* [2013]. Of course, the ensemble of TI realizations should be large enough

to ensure stable and thus reliable summary statistics.

To compress the model space, each TI realization is transformed into the discrete cosine domain and the dominant transform coefficients are determined (see Section 7.2.1). The DCT of the TI realizations are then truncated by setting all coefficients to zero except the dominant ones (note that these dominant coefficients are representative of all TI realizations). The inverse DCT of these truncated spectra yields an ensemble of porosity realizations. These porosity realizations, rather than their original counterparts, are used to extract summary metrics. In this way, we ensure that the summary statistics describe features that can be represented by the sparse model parameterizations. The various steps of this procedure are shown in Figure 7.1.

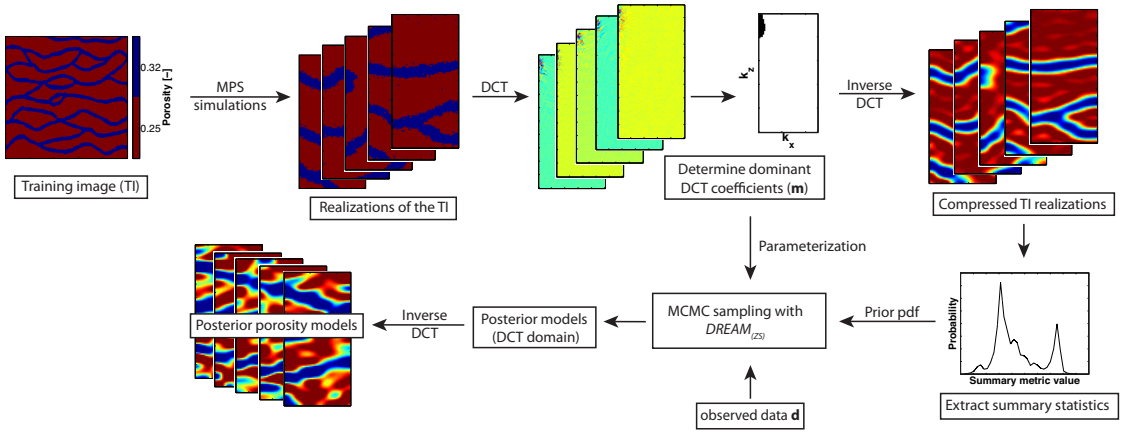


Figure 7.1: Schematic workflow of the presented inversion strategy. Realizations of a training image (TI) generated by multiple-point statistics (MPS) simulations are subject to discrete cosine transformation (DCT). Dominant DCT coefficients are derived and the corresponding basis functions are used to parameterize the porosity fields. From compressed TI realizations, summary statistics distributions are extracted and translated into a prior probability density function. The model space is sampled with the DREAM$_{(ZS)}$ algorithm to obtain posterior model realizations.

To illustrate our methodology, we use three different summary metrics of model morphology: (a) overall model roughness, (b) parameter variability, and (c) the porosity percentiles. These three metrics summarize important information about the expected spatial distribution of porosity. Furthermore, they are easy to calculate and invoke only minimal computational cost.

The roughness of the model, S_r, is calculated from the model representation on a Cartesian grid, $\mathbf{A}$, using first-order differences in the x- and z-direction

$$S_r = \sum_{x=2}^{N_x} \sum_{z=2}^{N_z} \left(|A(x,z) - A(x-1,z)| + |A(x,z) - A(x,z-1)| \right). \tag{7.12}$$

Such a roughness term is, in a least-squares sense, part of the objective function in smoothness-constrained deterministic inversions, where the goal is to find least-structured models that explain the data (e.g., *Constable et al.* [1987]). In these approaches any measure of structure larger than zero is penalized, thereby resulting in models that are often overly smooth. Here, we use the TI realizations to extract a distribution describing the expected roughness. A global measure of roughness similar to S_r is often imposed as a constraint in image processing, where it is commonly referred to as total variation (TV, e.g., *Rudin et al.* [1992]).

Another summary metric used herein is the total sum of the absolute values of DCT coefficients used to parameterize the model

$$S_v = \sum_{k_x=1}^{N_x} \sum_{k_z=1}^{N_z} |\mathbf{B}_{\text{trunc}}(k_x, k_z)|, \tag{7.13}$$

where $\mathbf{B}_{\text{trunc}}$ is the sparse model representation in the DCT domain. This metric measures the expected variability, or energy, within the model image as S_v is large for models with many interfaces and strongly varying porosities at different spatial scales. The TI realizations thus allow us to extract explicit information about the expected variability.

To make sure that our posterior models exhibit the "right" porosity distribution, we consider as third, and last summary metric namely the 10, 50 and 90% percentiles of the porosities from our TI realizations. We summarize these values in the vector $\mathbf{S}_p = (S_{10}, S_{50}, S_{90})$.

7.2.4 Prior distribution based on summary statistics

Our model vector contains the n dominant DCT coefficients that define the porosity field, a set of petrophysical parameters that relate the porosity to GPR wavespeed values, and the relative error level σ_{rel}. All model parameters are initially sampled from uniform distributions. The ranges of the retained DCT coefficients are set equal to their maximum and minimum values found in the ensemble of TI realizations. Further, the upper and lower bounds of the petrophysical parameters are based on literature values [*Keller*, 1988]. We assume that the relative error level follows a Jeffreys prior (i.e., its logarithm is uniformly distributed) with a range of half and two times its true value [*Tarantola*, 2005].

The premise of this chapter is that summary metrics from TI realizations can help to constrain model morphology. The idea is to translate the distribution of these summary metrics observed in the TI realizations into a prior distribution. This can thus be seen as an indirect prior on the model parameters, as the summary metrics of a model depend on the porosity values which in turn depend on the sampled DCT coefficients and petrophysical parameters.

To handle prior distributions of arbitrary shape, we create histograms of the summary metrics of the TI realizations. The frequencies on the y-axis are scaled so that the histogram integrates to unity, and represents a proper probability distribution. The prior probability of a proposal model, $p(\mathbf{m}_{\text{prop}})$, is in the case of independent summary statistics given by

$$p(\mathbf{m}_{\text{prop}}) = \prod_{j=1}^{d_S} p_{j,k}, \tag{7.14}$$

where d_S signifies the dimensionality of $\mathbf{S}$ and $p_{j,k}$ is the probability of the k-th bin for the j-th summary metric. The assumption of independence among the summary metrics is acceptable for the study considered herein (e.g., correlation coefficients smaller than 0.1 for the first case study) but might not be reasonable for other problems and a multivariate prior

distribution should be used.

This distribution constitutes a bounded indirect prior, as the bins are restricted to a certain range and the probability assigned to each bin refers to the summary metric, not to the sampled individual model parameters. This, however, does not guarantee that the summary metrics of the proposed models will fall into the appropriate range covered by this prior. For example, please consider Figure 7.2 that presents a histogram of the sampled model roughness values for one of the case studies discussed later. The gray bars depict the roughness distribution in an ensemble of 1,000 TI realizations. Each bin is assigned a probability p_k (black solid line). The roughness values of the models from the initial archive $\mathbf{Z}$ used in DREAM$_{(ZS)}$ are shown in blue. None of these models are deemed acceptable, as their roughness values fall outside the prescribed prior distribution for this summary metric. This happens as DREAM$_{(ZS)}$ in the initiation stage draws independent realizations of all model parameters (i.e., DCT coefficients and petrophysical parameters) without any consideration of the summary statistics.

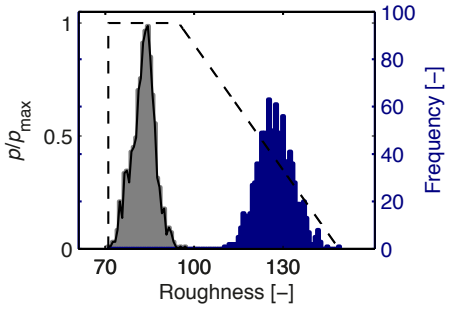


Figure 7.2: Formulation of the summary metrics prior and the indicator function. The distribution of roughness values observed in a set of 1,000 TI realizations is shown in gray and the estimated prior following this distribution is shown by the black solid line. The blue bars depict the distribution of roughness values in an initial set of random models. To ensure that the posterior models have roughness values within the summary metric prior range, an indicator function (dashed black line) is applied. The indicator function is 1 where the summary metrics prior is non-zero and it decreases to 0 towards the possible extreme values that might be sampled. Only realizations for which the indicator function is 1 are considered as possible realizations of the posterior.

To ensure that all models honor the observed summary metrics we draw inspiration from

Sadegh and Vrugt [2014] and use an indicator function on the prior distribution. This function $I(\mathbf{m})$ returns 1 if the summary metrics within a proposed model realization is within the range of the observed summary metrics derived from the TI realizations. Outside this range, $I(\mathbf{m})$ decreases linearly to 0 at the extreme ends. The acceptance rule to move from one model state to the next is then modified such that $\mathbf{m}^i_{\text{prop}}$ is always accepted if $I(\mathbf{m}^i_{\text{prop}}) > I(\mathbf{m}^i)$, otherwise the proposal is rejected unless $I(\mathbf{m}^i_{\text{prop}}) = 1$. In this latter case, the acceptance ratio is calculated following Equation 7.11. The range over which the indicator function is defined is determined by the minimum and maximum values of the summary metrics observed in the combined set of TI realizations and initial archive of models, $\mathbf{Z}$. If several summary metrics are considered, the indicator values are multiplied, meaning that all summary metric values must lie in the prior range for the proposal to be evaluated as a possible posterior sample.

As pointed out by *Sadegh and Vrugt* [2014], this modified acceptance rule introduces two different search stages of DREAM$_{(\text{ZS})}$. Initially, the algorithm acts as an optimizer and only accepts proposals if the summary metrics of the proposals are in better agreement with their observed values (smaller Euclidean distance). The resulting transitions of the Markov chain are irreversible, and hence the sampled states in stage 1 cannot be used to infer statistical properties of the model parameters. Once the simulated summary metrics are within their desired prior range $\left(I(\mathbf{m}^i_{\text{prop}}) = 1\right)$, stage 2 is initiated and Equation 7.11 is used to determine whether to accept each proposal or not. This then constitutes a regular MCMC step, and should, after sufficient burn-in, provide a sample of the posterior target distribution. In stage 2, it regularly happens that the proposal jump (Equation 7.9) leads to models with summary statistics that are outside of the desired prior range. When this happens, new proposals are generated with Equation 7.9 until $I(\mathbf{m}^i_{\text{prop}}) = 1$. The acceptance probability of Equation 7.11 then determines whether this proposal is accepted, otherwise the chain remains at its current state. Empirical results presented herein illustrate that this two-stage acceptance rule works well in practice.

The summary statistics have been constructed from a population of TI realizations and can be represented with our model parameterization. This guarantees the presence of a region within the model space which satisfies the measured values of the summary metrics. However, this does not discount the possibility of multiple disconnected behavioral regions that each honor the observed summary metrics. Such disconnected posterior distributions are very difficult to sample adequately. Fortunately, by using $K = 5$ different chains in DREAM$_{(ZS)}$ the model space is exhaustively explored, and by setting the jump rate periodically to unity in Equation 7.9, the chains can jump directly from one mode to the next. This enables inference of complex and multimodal posterior distributions [*Vrugt et al.*, 2009a, *Laloy and Vrugt*, 2012].

7.2.5 Forward modeling

The proposed inversion method is applied to first-arrival travel time data from synthetic crosshole GPR experiments. A radar wave is transmitted in one borehole and recorded in an adjacent one. By measuring the first-arrival time of the signal for various transmitter-receiver configurations, we can derive information about the spatial distribution of the radar slowness u (i.e., the reciprocal of the wavespeed) between the boreholes, which for a water-saturated porous medium can be related to the 2-D porosity distribution $\boldsymbol{\varphi}(x, z)$ through the following petrophysical relation *Pride* [1994], *Davis and Annan* [1989]

$$u(x, z) = \frac{1}{c}\sqrt{\boldsymbol{\varphi}(x, z)^m(\epsilon_w - \epsilon_s) + \epsilon_s}, \tag{7.15}$$

where c [m/s] signifies the speed of light, m [-] denotes the cementation factor, and ϵ_w [-] and ϵ_s [-] are the relative electrical permittivitivities of water and grains, respectively. In our analysis, we assume that $c = 3 \times 10^8$m/s and $\epsilon_w = 81$, and consider $m \in [1.3, 1.8]$ and $\epsilon_s \in [2, 6]$ to be free parameters in the inversion. The ranges of m and ϵ_s are taken quite

wide and are based on literature values [*Keller*, 1988]. We refer to *Peterson* [2001] for a detailed discussion about suitable field procedures when acquiring crosshole GPR data and an assessment of the sources of errors that affect the picked travel times.

The forward problem thus involves the calculation of the GPR travel times between the two boreholes, which is a function of the unknown porosity field and petrophysical relationship. We obtain the spatial distribution of the first-arrival travel times for each source position by solving the eikonal equation using a finite-difference scheme [*Podvin and Lecomte*, 1991].

7.3 Results

To benchmark our inversion method, we use two synthetic case studies. The first study involves a subsurface of channel structures within a homogeneous matrix. This results in a binary TI. The second study involves a more complex geological setting featuring five lithological units forming typical sedimentary structures of a fluvial deposit. Both cases are described in detail by *Lochbühler et al.* [2014b]. The TIs for both studies and a few of their corresponding TI realizations are shown in Figure 7.3. A total of 1,000 realizations were generated from each of the TIs and compressed to $n = 100$ and $n = 150$ DCT coefficients, respectively (see section 7.2.1). For both case studies we used standard settings of the algorithmic variables of DREAM$_{(ZS)}$ [*Laloy and Vrugt*, 2012]. Convergence was assessed using the $\hat{R}$-statistic of *Gelman and Rubin* [1992]. This statistic compares for each parameter of interest the between- and within-variance of the chains. Because of the asymptotic independence, the between-member variance and $\hat{R}$ can be estimated consistently from a single DREAM$_{(ZS)}$ trial. In practice, values of $\hat{R}$ smaller than 1.2 indicate convergence to a limiting distribution. In all our calculations reported herein we use the last 50% of the samples in each chain to calculate the $\hat{R}$ diagnostic. Simulation results show that about 40,000 and 70,000 iterations were required to reach convergence for case study 1 and 2, respectively.

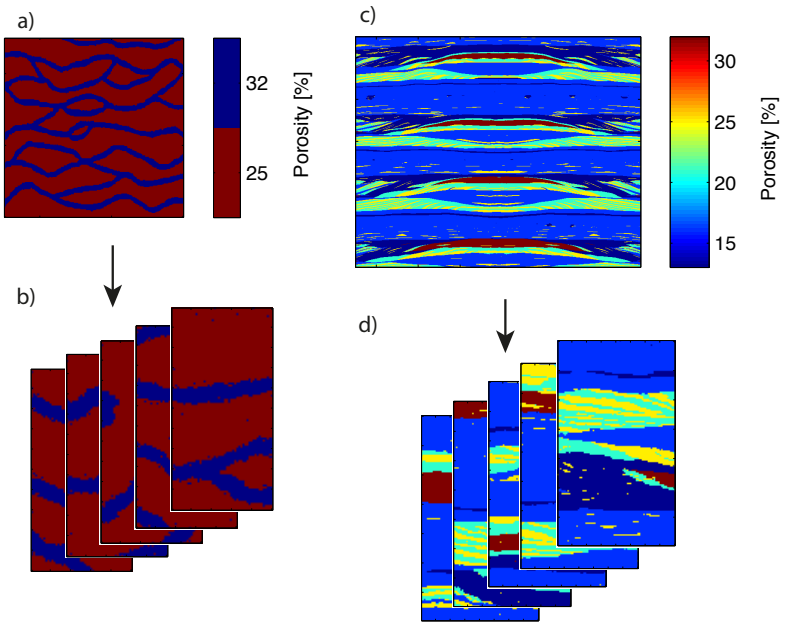


Figure 7.3: Training images for the channels case (a), the fluvial deposits case (c) and multiple-point statistics realizations thereof (b and d). Note that the full ensemble of TI realizations contains 1,000 realizations for both cases.

7.3.1 Channels case

Our first study, hereafter referred to as the channels case, considers channels of porosity $\varphi = 25\%$ that pervade a matrix with $\varphi = 32\%$. The relative electrical permittivity of the grains is $\epsilon_s = 3$ and the cementation factor, m, is 1.5. The forward problem is solved on a grid with square blocks of 0.1×0.1 m. GPR transmitter and receiver antennas are placed at a distance of 0.3 m along the left and right boundary of the domain (c.f. white dots in Figure 7.4a), resulting in a total of 879 travel times. The simulated travel times are corrupted with a heteroscedastic Gaussian measurement error with standard deviation set to 2% of the "observed" values. A relative error is used here to partly account for the typically larger observational and modeling errors associated with high-angle ray paths that are often associated with longer travel times, but similar results are expected if using absolute errors. The reference porosity field and its compressed image are shown in Figures 7.4(a) and (b), respectively. As we use a sparse model parameterization, the compressed image is, in a visual sense, the "best" model we can theoretically recover.

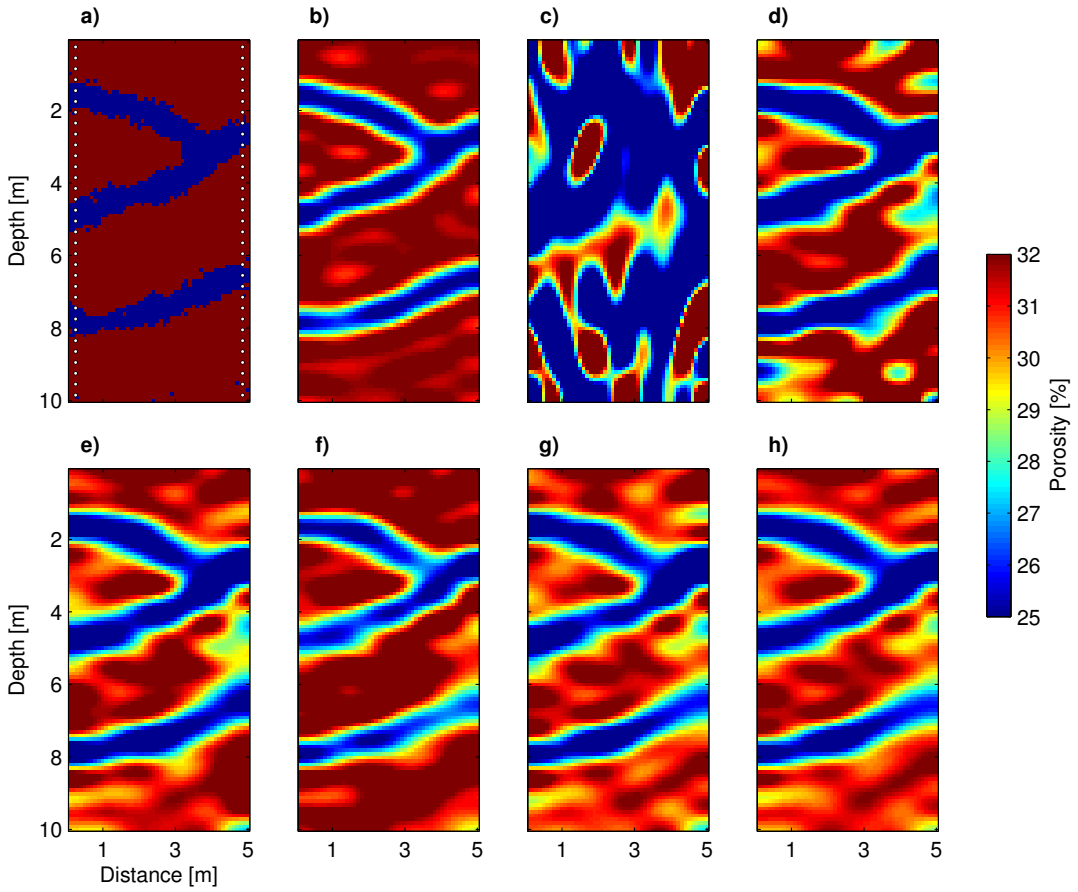


Figure 7.4: Channels case: a) reference porosity field with GPR transmitter (left side) and receiver (right side) antenna locations indicated by white dots, b) compressed image of the reference porosity field obtained by retaining the $n = 100$ highest amplitude DCT coefficients of the reference field. c) Ensemble mean of the summary metrics prior (variability and porosity percentiles) using the TI-based parameterization obtained by running an MCMC simulation with the likelihood set to unity. d) and e) Mean of the inferred porosity posterior without consideration of a summary metrics prior without (d) and with (e) TI-based parameterization. f)-i) Posterior mean of models obtained with TI-based parameterization and considering a summary metrics prior using f) roughness metric, g) variability metric, h) porosity percentile metrics and i) both variability and percentile metrics.

We first draw random simulations of our summary-statistics based prior using the TI-based parameterization by running the MCMC algorithm with a likelihood of unity. Eight sample realizations (Figure 7.5a) suggest that the summary-statistics based prior favors correct proportions between high- and low-porosity regions, but that the low-porosity regions (channel material) are generally disconnected. The ensemble mean (Figure 7.4c) has no resemblance with the reference model (Figures 7.4a-b), but the patterns are not truly random. This

is a consequence of our model parameterization strategy. It is clear that any channel-like structures that might appear in the posterior model realizations are the consequence of conditioning to the geophysical data. To illustrate the effect of the chosen model parameterization, the data were first inverted without a summary metrics-based prior distribution or a TI-based model parameterization. Instead, all but the 100 low-frequency coefficients arranged in a 10×10 rectangle were set to zero [*Linde and Vrugt*, 2013, *Lochbühler et al.*, 2014a]. Even though the data are almost fitted up to their measurement error, the inverse models bear very little resemblance with the reference porosity field (Figures 7.4d and 7.5b). With a TI-based model parameterization (section 7.2.1), the inversion results are markedly improved (Figures 7.4e and 7.5c) in that the main structural features of the reference porosity field are captured. However, inversion artifacts are abundant and m and ϵ_s are poorly estimated (Figures 7.5c, 7.6b). The use of a summary statistics-based prior distribution further improves the posterior models. Many of the inversion artifacts are removed and the petrophysical parameters are much closer to their true values (Figures 7.6c-f), particularly if the prior distribution is based on the observed sum of DCT coefficients (Figure 7.5e).

To quantify the quality of the posterior models, we calculated the root mean square error (RMSE) between the reference porosity field $\mathbf{A}_{\text{ref}}$ (c.f., Figure 7.4a) and the posterior porosity models, $\mathbf{A}$,

$$\text{RMSE} = \left|\left| \frac{1}{N_x N_z} (\mathbf{A}_{\text{ref}} - \mathbf{A}) \right|\right|_2, \tag{7.16}$$

As shown in Table 7.1, the posterior models provide better porosity estimates if a TI-based parameterization is applied and the results are further improved if summary statistics are used. The closest resemblance to the reference model is achieved by joint use of the variability and the percentile metrics (mean RMSE=2.27%).

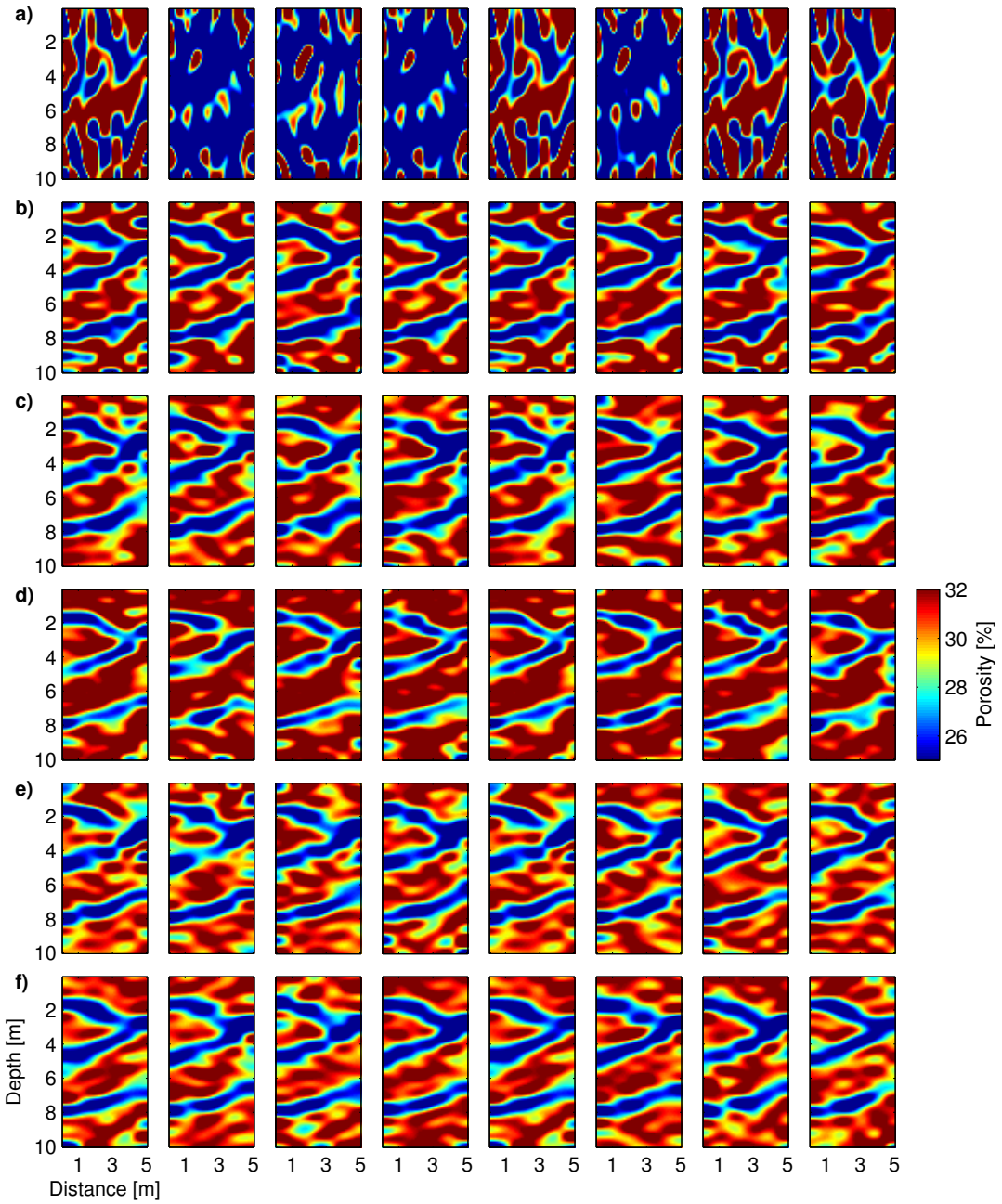


Figure 7.5: a) Random samples from the ensemble of summary metrics prior (variability and porosity metrics) models using the TI-based parameterization for the channels case. Corresponding random samples from the ensemble of posterior models for the channels case: b) no TI-based parameterization, no summary metric prior, c) TI-based parameterization, no summary metric prior, d)-g) TI-based parameterization, summary metric prior using d) roughness metric, e) variability metric, f) porosity percentile metrics, g) both variability and percentile metrics.

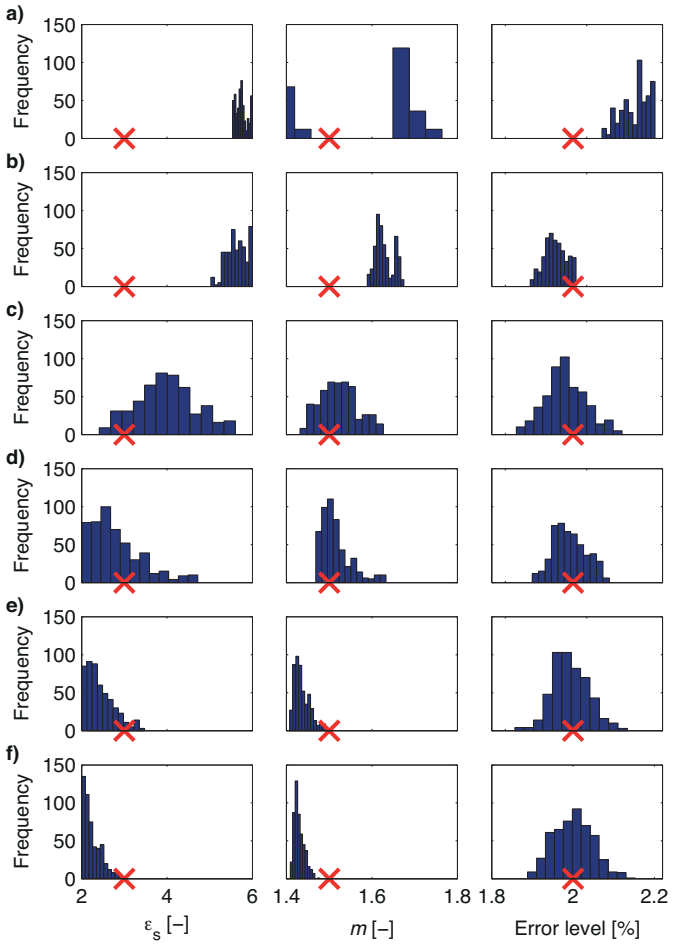


Figure 7.6: Posterior distributions of the petrophysical parameters ϵ_s and m, and of the relative data error level for the channels case: a) no TI-based parameterization, no summary metric prior, b) TI-based parameterization, no summary metric prior, c)-f) TI-based parameterization, summary metric prior using c) roughness metric, d) variability metric, e) percentile metrics, f) both variability and percentile metrics. Red crosses depict the reference values.

Table 7.1: Mean and standard deviation of the average root mean square error between posterior porosity models and the reference model for different degrees of prior information.

Considered prior information	mean RMSE [%]	std of RMSE [%]
Channels case		
no TI-based parameterization, no summary metric	9.37	0.84
TI-based parameterization, no summary metric	3.30	0.14
TI-based parameterization, roughness metric	2.57	0.10
TI-based parameterization, variability metric	2.36	0.10
TI-based parameterization, percentile metric	2.45	0.11
TI-based parameterization, variability and percentile metrics	2.27	0.10
Fluvial deposits case		
no TI-based parameterization, no summary metric	12.30	1.20
TI-based parameterization, no summary metric	5.44	0.63
TI-based parameterization, roughness and percentile metrics	4.20	0.31

The aberrant values of the petrophysical parameters demonstrate that it is not particularly easy to find "physically realistic" models in high-dimensional inversion. There are many models for which the data are adequately explained even if the petrophysical parameters are poorly estimated. This finding is perhaps not surprising, as the DCT parameterization is flexible enough to correct for this. The 2% measurement error is contained in the posterior distribution for all the different error level estimates (Figure 7.6, rightmost column), except for the inversion with no TI-based parameterization (Figure 7.6a).

To further demonstrate the benefits of the TI-based model parameterization and the summary metrics prior, we calculated the log posterior probability $\log(p(\mathbf{m}|\mathbf{d}))$, with m and ϵ_s varying between their prior bounds and the DCT coefficients kept fixed at their maximum a-posteriori density (MAP) values (Figure 7.7). The TI-based parameterization does not have a strong effect on the posterior density of the petrophysical parameters (compare Figures 7.7a-b), but the results are quite different when the summary statistics-based prior is used (Figures 7.7c-f). The MAP values (black crosses) are now in closer agreement with the m and ϵ_s values used to generate the synthetic GPR data (compare black and red crosses in Figure 7.7). The results not only favor the use of summary metrics, but also demonstrate the ability of the DREAM$_{(ZS)}$ algorithm, with modified acceptance rule, to locate and sample the underlying target distribution. Indeed, the MAP values of the petrophysical parameters (blue asterisks) coincide with the location of the posterior maximum (black crosses).

To provide more insights into the sampled posterior models, consider Figure 7.8 that plots prior and posterior distributions of the five different summary metrics used herein. The summary metrics of the posterior models lie exactly within their respective prior distribution (as expected), yet, they congregate in regions with relatively low prior probability (see Figures 7.8a and d). Nevertheless, the indicator function equals 1 (not shown), and the models are deemed behavioral. Note however, that none of the posterior models envelop the summary metrics of the reference model, except for the 90% percentile of the simulated porosity values

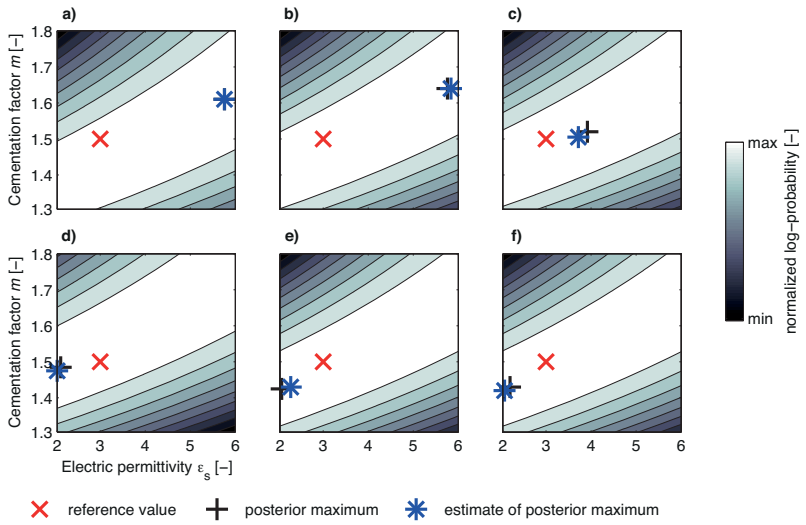


Figure 7.7: Maps of the posterior log-probability for all possible combinations of the petrophysical parameters m and ϵ_s for the channels case. The DCT coefficients are fixed at their maximum a-posteriori density (MAP) values. Indicated are the parameter values of the reference (red crosses), the values for which the log-probability is maximized (black crosses) and the estimate found by DREAM(ZS) (blue asterisks). Enumerations correspond to those of Figure 7.6.

(Figure 7.8e). This highlights the dominance of the likelihood term (many data) over the prior term (few summary statistics) in calculation of the posterior density.

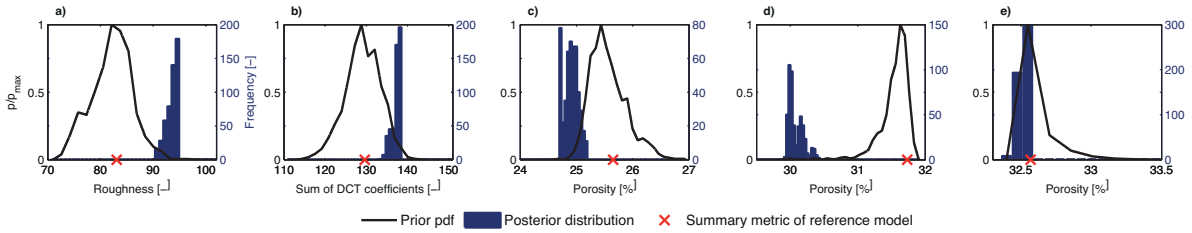


Figure 7.8: Summary metric prior distributions (black lines, see Figure 7.2 for explanation) and the posterior distributions of the summary metrics (blue bars) for the channels case: a) roughness metric, b) variability metric, c)-e) 10, 50, 90% percentile metrics. Red crosses display the summary metrics of the reference porosity field.

7.3.2 Fluvial deposits case

The second study, hereafter referred to as fluvial deposits case, is significantly more complex (Figures 7.9a-b). Five lithofacies with porosities ranging between 13 and 32% create a

sedimentary environment of fluvial deposits [*Bayer et al.*, 2011, *Comunian et al.*, 2011, *Lochbühler et al.*, 2014b]. We discretize the forward problems using blocks of 0.05×0.05 m and 0.25 m sensor spacing for the GPR experiments. This results in a total of 703 simulated travel times. The values of the petrophysical parameters ϵ_s and m and the relative magnitude of the data measurement error are equivalent to those used in the channels case.

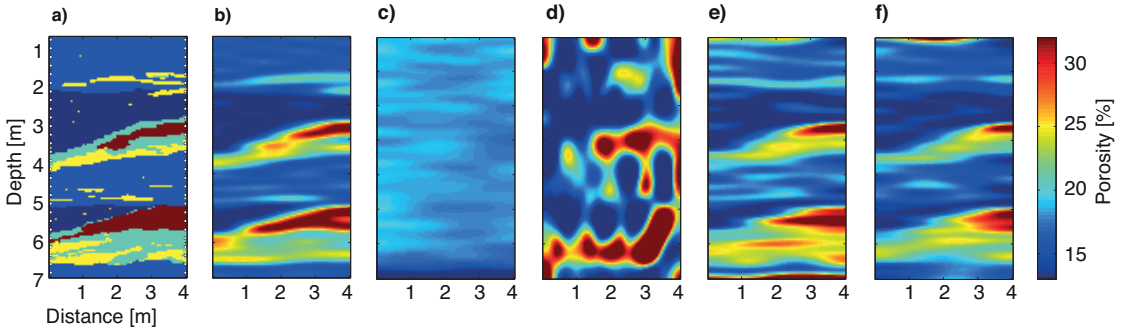


Figure 7.9: Fluvial deposits case: a) reference porosity field with GPR antenna locations indicated by white dots, b) compressed image of the reference porosity field obtained by retaining the $n = 150$ highest amplitude DCT coefficients of the reference field. c) Ensemble mean of the summary statistics prior (roughness and porosity percentiles) using the TI-based parameterization obtained by running an MCMC simulation with the likelihood always equal to 1. d) and e) Mean of the inferred porosity posterior without consideration of a summary metric prior without (d) and with (e) TI-based parameterization. f) Posterior mean of models with TI-based parameterization considering a summary metric prior based on roughness and porosity percentiles.

For this case we obtained the best results by combining the roughness and percentile metrics and we therefore restrict our analysis to those results. We first draw random simulations of our summary-statistics based prior using the TI-based parameterization by running the MCMC algorithm with a likelihood of unity. Eight sample realizations (Figure 7.10a) indicate that the summary-statistics based prior favors horizontal structures and a certain degree of variability. The ensemble mean (Figure 7.9c) shows little structure and provides no information about the actual location of high and low porosity regions. This information can only be obtained by considering the actual data. Inversion based on a model parameterization with a set of low-frequency DCT coefficients (here, a 15×10 rectangle is used) without a summary statistics-based prior does not produce adequate results (Figures 7.9d, 7.10b, 7.11a). The use of a TI-based parameterization leads to a much better recovery of the main structures of the reference field, and the added use of summary metrics further

improves the inversion results by suppressing inversion artifacts and improving the petrophysical parameter estimates (Figures 7.9e,f, 7.10c,d, 7.11b,c, Table 7.1). The relative error level is well retrieved in all cases that consider information from the TI realizations (Figures 7.11b-c) and the geophysical data. The benefit of using information from TI realizations is also illustrated in Figure 7.12. The MAP values are in much better agreement with their reference (true) values, particularly if summary metrics are used (Figures 7.12b-c).

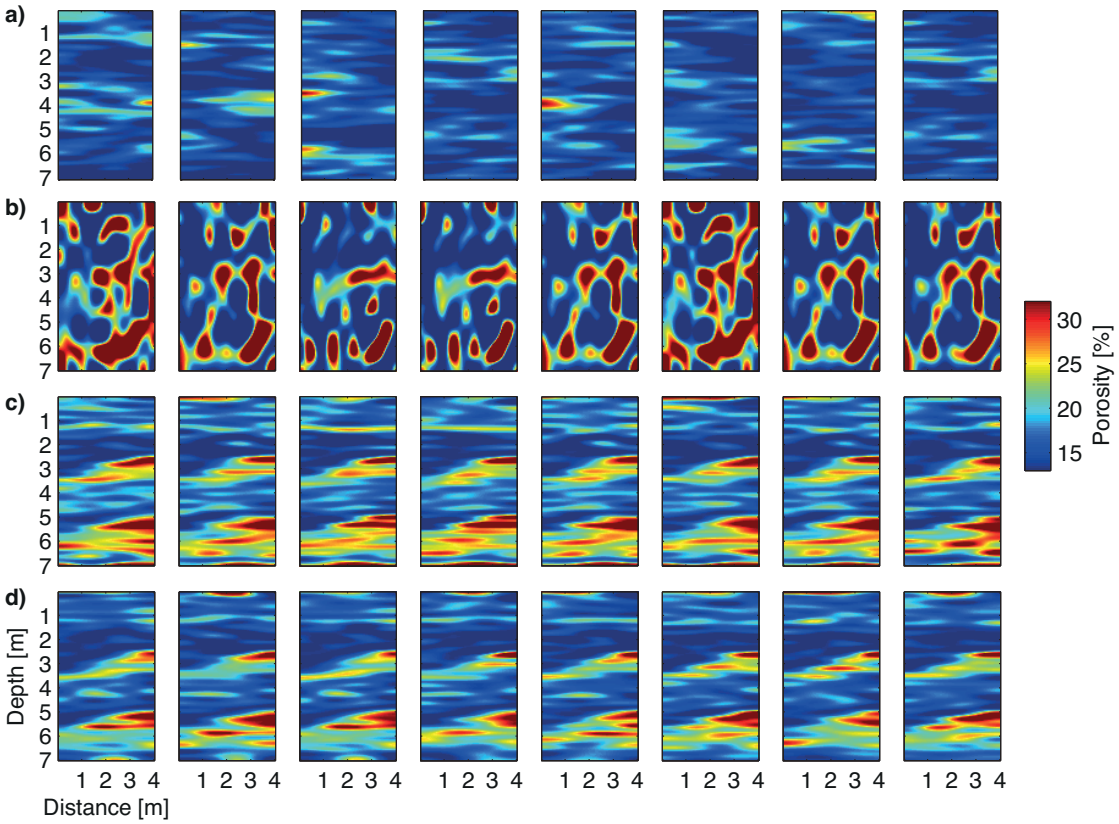


Figure 7.10: a) Random samples from the ensemble of summary statistics prior (roughness and porosity percentiles) using the TI-based parameterization for the fluvial deposits case. Corresponding random samples from the ensemble of posterior models: b) no TI-based parameterization, no summary metric prior, c) TI-based parameterization, no summary metric prior, d) TI-based parameterization, summary metric prior using roughness and porosity percentile metrics.

To provide further insights into the inversion results, consider Figure 7.13 that presents plots of the prior (black line) and posterior (blue histogram) distribution of each summary metric. The reference values of the summary metrics are separately indicated at the bottom of each plot with a red cross. The posterior estimates of the summary metrics are in close

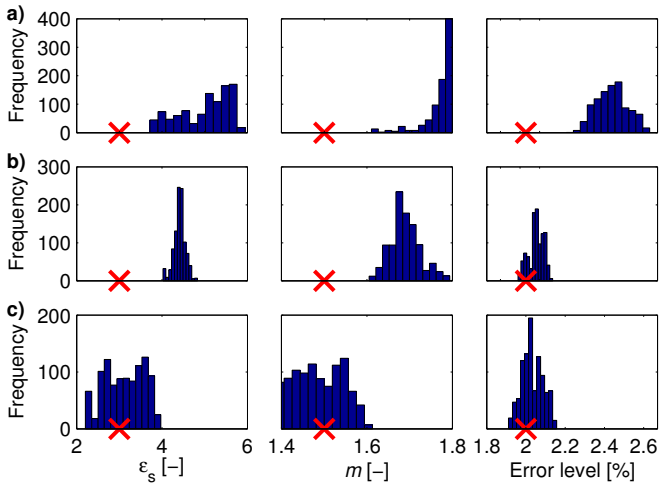


Figure 7.11: Posterior distributions of the petrophysical parameters ϵ_s and m, and of the relative data error level for the fluvial deposits case: a) no TI-based parameterization, no summary metric prior, b) TI-based parameterization, no summary metric prior, c) TI-based parameterization, summary metric prior using roughness and percentile metrics. Red crosses depict the reference values.

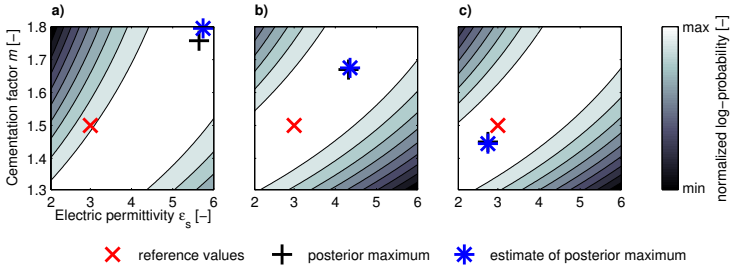


Figure 7.12: Maps of the posterior log-probability for all possible combinations of the petrophysical parameters m and ϵ_s for the fluvial deposits case. The DCT coefficients are fixed at their maximum a-posteriori density (MAP) values. Indicated are the parameter values of the reference (red crosses), the values for which the log-probability is maximized (black crosses) and the estimate found by DREAM$_{(ZS)}$ (blue asterisks). Enumerations correspond to those of Figure 7.11.

agreement with the corresponding values of the reference model. The percentiles of the porosity distribution are particularly well described and their marginal posterior distributions encompass the reference values.

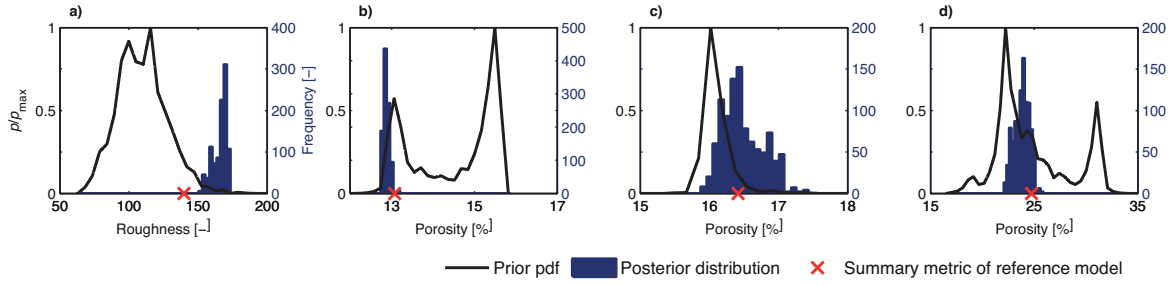


Figure 7.13: Prior probability density functions derived from the distribution of summary metrics in the ensemble of TI realizations (black lines, see Figure 7.2 for explanation) and the posterior distributions of the summary metrics (blue bars) for the fluvial deposits case: a) roughness metric, b)-d) 10, 50, 90% percentile metrics. Red crosses depict the summary metrics of the reference porosity field.

7.4 Discussion

We demonstrate that a TI can be used to define case-specific model parameterizations and to impose prior constraints on model morphology. This prior distribution is derived from an ensemble of TI realizations, and summarizes the frequency distribution of different summary statistics describing model morphology and lithology. The use of such prior restricts the feasible solution space, and significantly enhances the reliability of the posterior models. Applications are shown for crosshole radar tomography, but this general-purpose method can be applied to any type of data, as long as the observations are informative and sufficiently sensitive to the model parameters. In the two case studies considered herein, the summary metrics are extracted from 2-D MPS TIs, but the method is readily applicable to 3-D TIs. Furthermore, the TIs do not have to incorporate higher-order statistics. If the spatial distributions of the subsurface parameters are well-described by a multivariate Gaussian or other two-point statistical dependencies, and if these dependencies can be expressed as summary metrics, then it is straightforward to incorporate them in our methodology. The use of an indicator function enforces the posterior summary metrics to lie within the range derived from the TI realizations. This constitutes an optimization step before the posterior is reached [*Sadegh and Vrugt*, 2014], and hence violates detailed balance. As soon as the in-

dicator function is unity, and the simulated summary metrics lie within their respective prior distribution, the chain becomes Markovian as the indicator function is no longer informative, and the posterior distribution solely depends on the likelihood and prior probability.

The presented inversion framework can accommodate a wide range of summary statistics. The inversions presented herein are constrained by summary metrics describing the expected roughness, parameter variability, and percentile porosity distribution. The summary metrics used in each of the case studies significantly improved the inversion results, not only in terms of visual agreement with the reference model but also with respect to the posterior parameter estimates. Of course, the usefulness of each summary metric is case study dependent. The decision which summary metrics to use depends on the outcome of the analysis of the TI realizations. If, for instance, the prior distribution of a certain summary metric is flat then its use in the inversion would seem rather unproductive. On the contrary, metrics that are well-defined, are expected to help constrain the posterior models. For illustrative purposes, we have used rather simple and common metrics. If deemed appropriate, many other summary metrics can be defined and used, for example those that measure internal coherency [*McClymont et al.*, 2008], preferred lithological orientation (e.g., *Chugunova and Hu* [2008]), connectivity (e.g., *Renard et al.* [2011b]), sparseness (e.g., *Jafarpour et al.* [2009]), etc. More advanced summary metrics could be related to MPS specific information. Recently, *Lange et al.* [2012] constrained simulated annealing optimization with the pattern frequency from an MPS TI. *Honarkhah and Caers* [2010] used distance functions between pairs of patterns to quantify the characteristic statistics of a TI. Such metrics are easily incorporated in our inversion framework. We leave this for subsequent studies.

The use of the discrete cosine transform for model parameterization has the important advantage that the truncation level can be explicitly defined based on (a) the desired degree of small-scale structures that are to be resolved and (b) the sampling efficiency. The DREAM$_{(ZS)}$ algorithm is capable of handling a large number of parameters, but sampling

efficiency will decrease with increasing dimensionality of the search space. The truncation levels of $n = 100$ and $n = 150$ used in our case studies are sufficient to adequately represent the subsurface structures. The truncation drastically reduces the dimensionality of the model space compared to a Cartesian representation, which requires the estimation of 5,000 and 11,200 cells in the channels and the fluvial deposits case, respectively. Despite their sparseness, the use of a TI-based parameterization ensures that dominant structures are preserved in the posterior models. Our present formulation is based on a user-defined value n and it would be worthwhile to explore alternative formulations. Compressed sensing (e.g., *Candès et al.* [2006]), in which the sparsity of the parameter coefficients are maximized, or transdimensional (reversible-jump) MCMC (e.g., *Malinverno* [2002], *Bodin and Sambridge* [2009]), in which the number of model parameters are treated as an unknown throughout the inversion process, could be suitable starting points for developing a strategy to automatically estimate an appropriate value (or distribution) of n.

The main drawback of a DCT parameterization is that truncation impairs ability to accurately reproduce the sharp interfaces from the TI. Gradually updating a model according to an underlying geostatistical model (e.g., *Mariethoz et al.* [2010a], *Hansen et al.* [2012c]) can help to preserve sharp interfaces, but such an approach requires geostatistical resimulation between model states and has to deal with strongly correlated posterior model realizations [*Cordua et al.*, 2012, *Ruggeri et al.*, 2013].

This study reveals some of the well-known problems of probabilistic inversion in high dimensional parameter spaces. Poorly estimated model parameters are easily compensated for by other parameters that are well defined. This we observe in the interplay between DCT coefficients and the petrophysical parameters. An important problem is that the parameter values used to generate the data are not necessarily the most likely ones (Figures 7.7 and 7.12). In other words, conditioning only on geophysical data (e.g. travel times) can produce misleading results. We show that incorporation of auxiliary information through the use

of summary metrics from TI realizations significantly improves the posterior models and parameter values. The method can thus be seen as a strategy to handle high dimensionality in probabilistic inversion.

7.5 Conclusions

Training images are conceptual geological models of a certain site of interest. We use information from realizations of a TI to reduce the ambiguity inherent in probabilistic inversion of geophysical data by determining TI-based sparse model representations and by imposing morphological constraints on the posterior models. As a novelty, the model constraints are based on summarizing statistical metrics extracted from realizations of a TI. These are honored by formulating prior probability functions based on the distribution of summary metrics observed in the TI realizations. The summary metrics considered are global measures of model roughness, variability within the model and percentile distribution of porosity. Two different case studies were used to demonstrate the usefulness and applicability of the proposed inversion strategy. The use of summary statistics suppresses inversion artifacts, and provides parameter estimates that exhibit smaller posterior uncertainty and are in better agreement with their observed values. The methodology is computationally efficient, and designed to accommodate a wide range of geophysical forward problems and summary statistics. The summary metrics used in this study are perhaps rather simple, but significantly improved the reliability of the posterior parameter estimates and models. Application of more advanced summary metrics should hold even greater promise.

Chapter 8

Conclusions

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. In this thesis we introduced Approximate Bayesian Computation as an existing statistical vehicle for diagnostic model evaluation. This approach was introduced in the statistical literature about three decades ago [*Diggle and Gratton*, 1984] for cases when an explicit likelihood (objective) function cannot be justified. Such approaches, also referred to as ABC, use one or multiple (sufficient) statistics to estimate the posterior parameter distribution. The premise behind ABC is that $\boldsymbol{\theta}'$ should be a sample from the posterior distribution as long as the distance between the observed and simulated summary statistics is smaller than some small value, ϵ. In the limit of ϵ going to zero, the behavioral solution space should converge to the actual posterior distribution, pending the assumption that the chosen summary statistic(s) is (are) near sufficient [*Pritchard et al.*, 1999, *Vrugt and Sadegh*, 2013]. 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 in Chapter 2 of this thesis to introduce and illustrate the ABC framework [*Vrugt and Sadegh*, 2013].

This methodology, also, helps bridge the gap between informal and formal Bayesian approaches. Numerical simulations presented in Chapter 3 of this thesis have shown that, if each observation is treated as a summary variable, then the likelihood-free inference obtains very similar results as the limits of acceptability approach of GLUE. One issue deserves special attention, and that is that within the limit of acceptability framework, the value of ϵ needs to be taken much larger than what is deemed statistically adequate. Standard applications of likelihood-free inference define a solution to be behavioral if the chosen summary statistics are within a small distance of their observed counterparts. Yet for hydrologic systems with many calibration observations the probability of happening upon a simulation run that yields exactly the same data set as the one observed will be extremely small. The effective observation error remedies this problem, but the magnitude of this value is typically much larger than the theoretical value of ϵ to guarantee converge to the true posterior parameter distribution [*Sadegh and Vrugt*, 2013].

Successful application of the likelihood-free inference requires availability of an efficient sampling method that rapidly explores the space of behavioral models. Commonly used rejection sampling approaches adopt a boxcar kernel (0/1) to differentiate between behavioral ("1") and non-behavioral ("0") solutions, and use full-dimensional updating in pursuit of the posterior parameter distribution. This approach might work well for low-dimensional problems (e.g. $d \leq 10$) but is not particularly efficient in high-dimensional parameter spaces which require partial (subspace) sampling to rapidly locate posterior solutions.

In Chapter 4 of this thesis, we have introduced DREAM$_{(\mathrm{ABC})}$ to permit diagnostic inference of complex system models. This approach uses Metropolis-within-Gibbs simulation with DREAM [*Vrugt et al.*, 2008a, 2009a] to delineate the space of behavioral (posterior) models.

Three different case studies involving a simple 1-dimensional toy problem, a 20-dimensional mixture of bivariate distributions, and a 14-dimensional hydrologic model calibration problem illustrate that DREAM$_{(\mathrm{ABC})}$ is about 3 - 1,000 times more efficient than commonly used ABC sampling approaches. This gain in sampling efficiency increases with dimensionality of the parameter space [*Sadegh and Vrugt*, 2014].

The diagnostic power of the proposed process-based model evaluation approach is shown in Chapter 5 of this thesis to help detect and diagnose the reasons of catchment nonstationarity. Our approach combines the strengths of diagnostic analysis, approximate Bayesian computation, and temporal analysis of watershed signatures (aka summary statistics) of the rainfall-runoff response to explain observed trends and/or shifts in discharge dynamics. Failure of a model to simulate accurately the observed temporal dynamics of the watersheds rainfall-runoff signatures is an indication of basin nonstationarity, and hence suggests that the catchment properties have changed during the time period under consideration. Indeed, a single model parameterization cannot adequately simulate temporal trends in the summary metrics of catchment response, and hence application of ABC cannot provide a behavioral model space [*Sadegh et al.*, 2015a].

Chapter 6 of this thesis introduces a new class of closed-form formulations for characterizing the information content of the flow duration curves, as an important watershed signature. Application of the proposed soil water retention functions (WRFs) to 408 watersheds of the MOPEX data set demonstrates that the van Genuchten and Kosugi models closely match the empirical FDCs of arid, semi-arid, and wet catchments. The proposed parametric expressions exhibit a superior performance over commonly used FDC models in the hydrologic literature. If quality of fit is of paramount importance then the 5-parameter Kosugi model is preferred as it provides the closest fit to the empirical FDCs. What is more, this mixture formulation of two WRFs can help to discern different stable states of the catchment. If the fitting coefficients of the FDC model are to be used for regionalization or as summary statistics in

diagnostic analysis then the 2-parameter formulation of Kosugi is preferred. The parameters of this model exhibit a negligibly small correlation and are therefore suitable metrics that are well defined by calibration against the observed FDCs [*Sadegh et al.*, 2015b].

The ABC framework and using summary statistics for extracting information for model-data synthesis hold great promises in different fields of study. Chapter 7 of this thesis demonstrates the application of the ABC framework for geophysical modeling. We characterize model constraints, in this chapter, based on summarizing statistical metrics extracted from realizations of a training image (conceptual geological models of a certain site of interest). These are honored by formulating prior probability functions based on the distribution of summary metrics observed in the TI realizations. The summary metrics considered are global measures of model roughness, variability within the model and percentile distribution of porosity. Two different case studies were used to demonstrate the usefulness and applicability of the proposed inversion strategy. The use of summary statistics suppresses inversion artifacts, and provides parameter estimates that exhibit smaller posterior uncertainty and are in better agreement with their observed values. The methodology is computationally efficient, and designed to accommodate a wide range of geophysical forward problems and summary statistics. The summary metrics used in this study are perhaps rather simple, but significantly improved the reliability of the posterior parameter estimates and models. Application of more advanced summary metrics should hold even greater promise [*Lochbühler et al.*, 2014c].

Altogether, 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. The ABC methodology opens an entire new field of research with infinite scope for (amongst others) (a) inventing new summary metrics that are self-sufficient and properly rooted in hydrologic and information theory, (b) developing methods and guidelines to interpret the outcome of the ABC analysis and help aid in detec-

tion of model malfunctioning [*Vrugt and Sadegh*, 2013], (c) exploring ways to determine the temporal/spatial variability of individual summary metrics for stochastic (Bayesian) analysis, (d) new approaches to diagnostic model selection using stochastic minimization of the mutual information, (e) improving the computational efficiency of ABC-sampling [*Sadegh and Vrugt*, 2013], and (f) practical applications. Challenges lie in the proper selection of summary metrics that properly extract the information from the calibration data, how to deal with input data uncertainty, and how to detect epistemic errors (lack of knowledge).

Bibliography

Ahmed, N., T. Natarajan, and K.R. Rao (1974), Discrete cosine transform, *IEEE Transactions on Computers*, *100*(1), 9093.

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 Resources Research*, *43*, W01403, doi:10.1029/2005WR004745.

Alaouze, C.M. (1989), Reservoir releases to uses with different reliability requirements, *Water Resources Bulletin*, *25*(6),1163–1168.

Alexander, L. V., et al. (2006), Global observed changes in daily climate extremes of temperature and precipitation, *Journal of Geophysical Research*, *111*, D05109, doi:10.1029/2005JD006290.

Anghileri, D., F. Pianosi, and R. Soncini-Sessa (2014), Trend detection in seasonal data: from hydrology to water resources, *Journal of Hydrology*, *511*, 171–179.

Arnold, J.G., and P.M. Allen (1999), Automated methods for estimating baseow and ground water recharge from streamow records, *Journal of the American Water Resources Association*, *35*, 411–424.

Aronica, G., P.D. Bates, and M.S. Horritt (2002), Assessing the uncertainty in distributed model predictions using observed binary pattern information within GLUE, *Hydrological Proccesses*, *16*, 2001–2016.

Assouline, S., D. Tessier, and A. Bruand (1998), A conceptual model of the soil water retention curve, *Water Resour. Res.*, *34*(2), 223-231.

Backus, G., and F. Gilbert (1970), Uniqueness in the inversion of inaccurate gross earth data. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *266*, 123192.

Barnes, C., S. Filippi, M.P.H. Stumpf, and T. Thorne (2011), Considerate approaches to achieving sufficiency for ABC model selection, *http://arxiv.org/pdf/1106.6281v2.pdf*, *1*(21).

Bassiouni, M., and D. S. Oki (2013), Trends and shifts in streamflow in Hawaii, 1913-2008, *Hydrological Processes*, *27*(10), 1484-1500.

Bates, B.C., and E.P. Campbell (2001), A Markov chain Monte Carlo scheme for parameter estimation and inference in conceptual rainfall-runoff modeling, *Water Resources Research*, *37*(4), 937–947.

Bayer, P., P. Huggenberger, P. Renard, and A. Comunian (2011), Three-dimensional high resolution fluvio-glacial aquifer analog: Part 1: Field study, *Journal of Hydrology*, *405*, 1–9.

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

Beaumont, M.A., J.M. Cornuet, J.M., Marin, and C.P. Robert (2009), Adaptive approximate Bayesian computation, *Biometrika*, *asp052*, 1–8.

Beaumont, M.A. (2010), Approximate Bayesian Computation in Evolution and Ecology, *Annual Review of Ecology, Evolution, and Systematics*, *41*, 379–406.

Bertorelle, G., A. Benazzo, and S. Mona (2010), ABC as a flexible framework to estimate

demography over space and time: some cons, many pros, *Molecular Ecology*, *19*, 2609–2625.

Beven, K.J., and A.M. Binley (1992), The future of distributed models: Model calibration and uncertainty prediction, *Hydrological Processes*, *6*, 279–298.

Beven, K.J. (1993), Prophecy, reality and uncertainty in distributed hydrological modeling, *Advances in Water Resources*, *16*(1), 41–51.

Beven, K., and J. Freer (2001), Equifinality, data assimilation, and uncertainty estimation in mechanistic modeling of complex environmental systems using the GLUE methodology, *Journal of Hydrology*, *249*, 11–29.

Beven, K., and P. Young (2003), Comment on "Bayesian recursive parameter estimation for hydrologic models" by M. Thiemann, M. Trosset, H. Gupta, and S. Sorooshian, *Water Resources Research*, *39*(5), 1116, doi:10.1029/2001WR001183.

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

Beven, K., P.J. Smith, and J. Freer (2008a), So why would a modeller choose to be incoherent?, *Journal of Hydrology*, *354*, 15–32.

Beven, K. (2009), Comment on "Equifinality of formal.(DREAM) and informal.(GLUE) Bayesian approaches in hydrologic modeling?" by Vrugt J.A, ter Braak C.J.F, Gupta H.V., and Robinson B.A., *Stochastic Environmental Research and Risk Assessment*, *23*(7), 1059–1060, doi:10.1007/s00477-008-0283-x.

Beven, K., P. Smith, and A. Wood (2011), On the colour and spin of epistemic error (and what we might do about it), *Hydrology and Earth System Sciences*, *15*, 3123–3133.

Bikowski, J., J.A. Huisman, J.A., Vrugt, H. Vereeken, and J. van der Kruk (2012), Inversion and sensitivity analysis of ground penetrating radar data with waveguide dispersion using

deterministic and Markov chain Monte Carlo methods, *Near Surface Geophysics*, Special issue "Physics-based integrated characterization", *xx*, xxx–xxx.

Birsan, M. V., P. Molnar, P. Burlando, M. Pfaundler (2005), Streamflow trends in Switzerland, *Journal of Hydrology*, *314*, 312-329.

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

Blum, M.G.B., and O. François (2010), Non-linear regression models for approximate Bayesian computation, *Statistics and Computing*, *20*, 63–73.

Bodin, T., and M. Sambridge (2009), Seismic tomography with the reversible jump algorithm, *Geophysical Journal International*, *178*(3), 1411–1436.

Booker, D.J., and T.H. Snelder (2012), Comparing methods for estimating flow duration curves at ungauged sites, *Journal of Hydrology*, *434-435*, 78–94.

Bosch, M., R. Meza, R. Jiménez, and A. Hönig (2006), Joint gravity and magnetic inversion in 3D using Monte Carlo methods. Geophysics, 71(4):G153G156.

Botter, G., S. Zanardo, A. Porporato, I. Rodriguez-Iturbe, A. and Rinaldo (2008), Ecohydrological model of flow duration curves and annual minima, *Water Resources Research*, *44*, W08418, doi:10.1029/2008WR006814.

Brazier R.E, K.J. Beven, S.G. Anthony, and J.S. Rowan (2001), Implications of model uncertainty for the mapping of hillslope-scale soil erosion predictions, *Earth Surface Processes and Landforms*, *26*, 1333–1352.

Brooks, R.H., and A.T. Corey (1964), Hydraulic properties of porous media, Hydrol. Pap. 3, Colo. State Univ., Fort Collins.

Buishand, T.A. (1982), Some methods for testing the homogeneity of rainfall records, *Journal of Hydrology*, *58*, 11–27.

Buland, A., and O. Kolbjørnsen (2012), Bayesian inversion of CSEM and magnetotelluric data, *Geophysics*, *77*(1), 33–42.

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

Burke, E.J., R.H.J. Perry, S.J. Brown (2010), An extreme value analysis of UK drought and projections of change in the future, *Journal of Hydrology*, *388*(131), doi:10.1016/j.jhydrol.2010.04.035.

Burn, D. H., M. A. Hag Elnur (2002), Detection of hydrologic trends and variability, *Journal of Hydrology*, *255*, 107-122.

Burnash, R.J., R.L. Ferral, R.L., and R.A. McGuire (1973), A generalized streamflow simulation system: Conceptual modeling for digital computers, Joint Federal-State River Forecast Center, Sacramento, CA, USA.

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, *Journal of Hydrology*, *298*, 242–266, doi:10.1016/j.jhydrol.2004.03.042.

Caers, J., and T. Zhang (2004), Multiple-point geostatistics: a quantitative vehicle for integrating geologic analogs into multiple reservoir models, *AAPG Memoir*, *80*, 383–394.

Candès, E.J., J.K. Romber, and T. Tao (2006), Stable signal recovery from incomplete and inaccurate measurements, *Communications on Pure and Applied Mathematics*, *59*, 1207–1223.

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, *Hydrology and Earth System Sciences*, *15*, 3411–3430, doi:10.5194/hess-15-3411-2011.

Castellarin, A., G. Galeati, L. Brandimarte, A. Montanari, and A. Brath (2004a), Regional flow-duration curves: reliability for ungauged basins, *Advances in Water Resources*, *27*, 953–965.

Castellarin, A., R.M. Vogel, and A. Brath (2004b), A stochastic index flow model of flow duration curves, *Water Resources Research*, *40*, W03104, doi:10.1029/2003WR002524.

Castellarin, A., G. Camorani, and A. Brath (2007), Predicting annual and long-term flow-duration curves in ungauged basins, *Adv. Water Resour.*, *30*, 937–953, doi:10.1016/j.advwatres.2006.08.006.

Chebana, F., T. B. M. J. Ouarda, and T. C. Duong (2013), Testing for multivariate trends in hydrologic frequency analysis, *Journal of Hydrology*, *486*, 519–530.

Chen, M.-H., Q.-M. Shao, and J.G. Ibrahim (2000), *Monte Carlo methods in Bayesian computation*, Springer New York.

Cheng, L., M. Yaeger, A. Viglione, E. Coopersmith, S. Ye, and M. Sivapalan (2012), Exploring the physical controls of regional patterns of flow duration curves Part 1: Insights from statistical analyses, *Hydrol. Earth Syst. Sci.*, *16*, 4435-4446, doi:10.5194/hess-16-4435-2012.

Cong, Z., D. Yang, B. Gao, H. Yang, and H. Hu (2009), Hydrological trend analysis in the Yellow River basin using a distributed hydrological model, *Water Resources Research*, *45*(7), W00A13, doi:10.1029/2008WR006852.

Chow, V.T. (1964), Handbook of applied hydrology, Mc-Graw Hill BookCo., New York, N.Y.

Christensen, S. (2004), A synthetic groundwater modeling study of the accuracy of GLUE uncertainty intervals, *Nordic Hydrology*, *35*, 45–59.

Chu, W., X. Gao, S. and Sorooshian (2010), Improving the shuffled complex evolution scheme for optimization of complex nonlinear hydrological systems: Application to the calibration of the Sacramento soil-moisture accounting model, *Water Resources Research*, *46*, W09530, doi:10.1029/2010WR009224, 2010.

Chugunova, T., and L. Hu (2008), Multiple-point simulations constrained by continuous auxiliary data, *Mathematical Geosciences*, *40*(2), 133–146.

Cigizoglu, H.K., and M. Bayazit (2000), A generalized seasonal model for flow duration curve, *Hydrol. Process.*, *14*, 1053–1067.

Clark, M.P., and A.G. Slater (2006), Probabilistic quantitative precipitation estimation in complex terrain, *Journal of Hydrometeorology*, *7*(1), 3–22.

Clark, M.P., A.G. Slater, D.E. Rupp, R.A. Woods, J.A. Vrugt, H.V. Gupta, T. Wagener, and L.E. Hay (2008), Framework for Understanding Structural Errors(FUSE): A modular framework to diagnose differences between hydrological models, *Water Resources Research*, *44*, W00B02, doi:10.1029/2007WR006735.

Clark, M.P., and D. Kavetski (2010), Ancient numerical daemons of conceptual hydrological modeling: 1. Fidelity and efficiency of time stepping schemes, *Water Resources Research*, *46*, W10510, doi:10.1029/2009WR008894.

Clarke, R. T. (2002), Estimating trends in data from the Weibull and a generalized extreme value distribution, *Water Resources Research*, *38*(6), doi:10.1029/2001WR000575.

Clarke, R. T. (2007), Hydrological prediction in a non-stationary world, *Hydrology and Earth System Sciences*, *11*, 408–414, doi:10.5194/hess-11-408-2007.

Clausen B., B.J.F. Biggs (2000), Flow variables for ecological studies in temperate streams: grouping based on covariance, *Journal of Hydrology*, *237(3-4)*, 184–197.

Cole, R.A.J., H.T. Johnston, and D.J. Robinson (2003), The use of flow duration curves as a data quality tool, *Hydrological Sciences Journal*, *48*(6), 939–951.

Comunian, A., P. Renard, J. Straubhaar, and P. Bayer (2011), Three-dimensional high resolution fluvio-glacial aquifer analog: Part 2: Geostatistical modeling, *Journal of Hydrology*, *405*, 10–23.

Constable, S., R. Parker, and C. Constable (1987), Occams inversion: A practical algorithm for generating smooth models from electromagnetic sounding data, *Geophysics*, *52*(3), 289–300.

Coopersmith, E., M.A. Yaeger, S. Ye, L. Cheng, and M. Sivapalan (2012), Exploring the physical controls of regional patterns of flow duration curves Part 3: A catchment classification system based on regime curve indicators, *Hydrol. Earth Syst. Sci.*, *16*, 4467-4482, doi:10.5194/hess-16-4467-2012.

Cordua, K., T. Hansen, and K. Mosegaard (2012), Monte Carlo full-waveform inversion of crosshole GPR data using multiple-point geostatistical a priori information, *Geophysics*, *77*(2), 19–31.

Croker, K.M., A.R. Young, M.D. Zaidman, and H.G. Rees (2003), Flow duration curve estimation in ephemeral catchments in Portugal, *Hydrological Sciences Journal*, *48*(3), 427–439, DOI: 10.1623/hysj.48.3.427.45287.

Csilléry, K., M.G.B. Blum, O.E. Gaggiotti, and O. François (2010), Approximate Bayesian Computation (ABC) in practice, *Trends in Ecology & Evolution*, *25*, 410–418.

Cunnane, C. (1988), Methods and merits of regional flood frequency analysis, *Journal of Hydrology*, *100*(1-3), 269–290.

Cunderlik, J. M., and D. H. Burn (2003), Non-stationary pooled flood frequency analysis, *Journal of Hydrology*, *276*, 210-223.

Cunderlik, J. M., and T. B. M. J. Ouarda (2006), Regional flood-duration-frequency modeling in a changing environment, *Journal of Hydrology*, *318*, 276–291.

Davis, J., and A. Annan (1989), Ground-penetrating radar for high-resolution mapping of soil and rock stratigraphy, *Geophysical Prospecting*, *37*(5), 531–551.

Davis, K., and Y. Li (2011), Fast solution of geophysical inversion using adaptive mesh, space-filling curves and wavelet compression. *Geophysical Journal International*, *185*(1), 157–166.

DeChant, C. M., and H. Moradkhani (2012), Examining the effectiveness and robustness of sequential data assimilation methods for quantification of uncertainty in hydrologic forecasting, *Water Resources Research*, *48*, W04518, doi:10.1029/2011WR011011.

Del Moral, P., A. Doucet, and A. Jasra (2008), An adaptive sequential Monte Carlo method for approximate Bayesian computation, Technical Report, Imperial College London.

Dette, H., A. Munk, and T. Wagner (1998), Estimating the variance in nonparametric regression: What is a reasonable choice?, *Journal of the Royal Statistical Society Series B, Part 4, 60*, 751–764.

Dettinger, M. (2011), Climate change, atmospheric rivers, and floods in California a multimodel analysis of storm frequency and magnitude changes. *JAWRA Journal of the American Water Resources Association*, *47*, 514-523.

Dhanya, C.T. and D.N. Kumar (2011), Predictive uncertainty of chaotic daily streamflow using ensemble wavelet networks approach, *Water Resources Research*, *47*, W06507, doi:10.1029/2010WR010173.

Didelot, X., R.G. Everitt, A.M. Johansen, and D.L. Lawson (2011), Likelihood-free estimation of model evidence, *Bayesian Analysis*, *6* (1), 49–76.

Diggle, P.J., and R.J. Gratton (1984), Monte Carlo methods of inference for implicit statistical models, *Journal of the Royal Statistical Society Series B*, *46*, 193–227.

Dingman, S.L. (1978), Sythesis of Flow-Duration Curves for Unregulated Streams in New Hampshire, *Water Resources Boulletin*, *14*(6), 1481–1502.

Doherty, J. and D. Welter (2010), A short exploration of structural noise, *Water Resources Research*, *46*, W05525, doi:10.1029/2009WR008377.

Douglas, E. M., R. M. Vogel, and C. N. Kroll (2000), Trends in floods and low flows in the United States: impact of spatial correlation, *Journal of Hydrology*, *240*(1-2), 90–105.

Duan, Q., S. Sorooshian, and V. Gupta (1992), Effective and efficient global optimization for conceptual rainfall-runoff models, *Water Resources Research*, *28*(4), 1015–1031.

Duan, Q., S. Sorooshian, V.K. Gupta (1994), Optimal use of the SCE-UA global optimization method for calibrating watershed models, *Journal of Hydrology*, *158*(3-4), 265–284.

Duan, Q., J. J. Schaake, V. Andréassian, S. Franks, G. Goteti, H.V. Gupta, Y.M. Gusev, F. Habets, A. Hall, L. Hay, T. Hogue, M. Huang, G. Leavesley, X. Liang, O.N. Nasonova, J. Noilhan, L. Oudin, S. Sorooshian, T. Wagener, and E.F. Wood (2006), Model parameter estimation experiment (MOPEX): An overview of science strategy and major results from the second and third workshops, *Journal of Hydrology*, *320*, 3–17.

Durner, W. (1994), Hydraulic conductivity estimation for soils with heterogeneous pore structure, *Water Resources Research*, *30*(2), 211–223.

Eckhardt, K. (2005), How to construct recursive digital lters for baseow separation, *Hydrological Processes*, *19*, 507–515.

Edwards, A. W. F. (1992), Likelihood, *The Johns Hopkins University Press*, Baltimore and London.

El-Adlouni, S., T. B. M. J. Ouarda, X. Zhang, R. Roy, and B. Bobée (2007), Generalized maximum likelihood estimators for the nonstationary generalized extreme value model, *Water Resources Research*, *43*, W03410, doi:10.1029/2005WR004545.

Euser, T., H. C., Winsemius, M. Hrachowitz, F. Fenicia, S. Uhlenbrook, and H. H. G., Savenije (2013), A framework to assess the realism of model structures using hydrological signatures, *Hydrology and Earth System Sciences*, *17*, 1893–1912, doi:10.5194/hess-17-1893-2013.

Evin, G., D. Kavetski, M. Thyer, and G. Kuczera (2013), Pitfalls and improvements in the joint inference of heteroscedasticity and autocorrelation in hydrological model calibration, *Water Resources Research*, *49*, doi:10.1002/wrcr.20284.

Fearnhead, P., and D. Prangle (2012), Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation, *Journal of the Royal Statistical Society Series B*, *74*, Part 3, 1–28.

Fennessey, N., and R.M. Vogel (1990), Regional Flow-Duration Curves for Ungauged Sites in Massachusetts, *J. Water Resour. Plann. Manage.*, *116*, 530–549.

Fenicia, F., H.H.G. Savenije, P. Matgen, and L. Pfister (2007), A comparison of alternative multiobjective calibration strategies for hydrological modeling, *Water Resources Research*, *43*, W03434, doi:10.1029/2006WR005098.

Feyen L, K.J. Beven, F. De Smedt, and J.E. Freer (1002), Stochastic capture zone delineation within the generalized likelihood uncertainty estimation methodology: conditioning on head observations, *Water Resources Research*, *37*(3), 625–638.

Foster, H.A (1934) Duration curves, *ASCE Transactions*, *99*, 1213–1267.

Franchini, M. and M. Suppo (1996), Regional analysis of flow duration curves for a limestone region, *Water Resour. Manage.*, *10*, 199-218.

Franks S.W., K.J. Beven, P.F. Quinn, and I.R., Wright (1997), On the sensitivity of soil-vegetation-atmosphere transfer(SVAT) schemes: equifinality and the problem of robust calibration, *Agricultural and Forest Meteorology*, *86*, 63–75.

Freer, J., K. Beven, and B. Ambroise (1996), Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach, *Water Resources Research*, *32*(7), 2161–2173.

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 Resources Research*, *47*, W12529, doi:10.1029/2010WR009993.

Fu, G., S. Chen, C. Liu, and D. Shepard (2004), Hydro-climatic trends of the Yellow River basin for the last 50 years, *Climate Change*, *65*(1-2), 149–178.

Gallardo, L., and M. Meju (2004), Joint two-dimensional DC resistivity and seismic travel time inversion with cross-gradients constraints, *Journal of Geophysical Research*, *109*(B3), B03311.

Ganora, D., P. Claps, F. Laio, and A. Viglione (2009), An approach to estimate non-parametric flow duration curves in ungauged basins, *Water Resour. Res.*, *45*, W10418, doi:10.1029/2008WR007472.

Gelman, A., D.B. Rubin (1992), Inference from iterative simulation using multiple sequences, *Statistical Science*, *7*, 457–472.

Georgakakos, K.P., D.J. Seo, H. Gupta, J. Schaake, and M.B. Butts (2004), Towards the char-

acterization of streamflow simulation uncertainty through multimodel ensembles, *Journal of Hydrology*, *298*, 222–241.

González, E.F., T. Mukerji, and G. Mavko (2007), Seismic inversion combining rock physics and multiple-point geostatistics, *Geophysics*, *73*(1), R11–R21.

Graf, W. L. (1977), Network characteristics in suburbanizing streams, *Water Resources Research*, *13*(2), 459-463.

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

Groisman, P. Y., R. W. Knight, and T. R. Karl (2001), Heavy precipitation and high streamflow in the contiguous United States: trends in the twentieth century, *Bulletin of American Meteorological Society*, *82*, 219-246.

Groisman, P.Y., R. W. Knight, T. R. Karl, D. R. Easterling, B. Sun, and J. H. Lawrimore (2004), Contemporary changes of the hydrological cycle over the contiguous United States: Trends derived from in situ observations. *Journal of Hydrometeorology*, *5*, 64-85.

Gupta, V.K. and S. Sorooshian (1983), Uniqueness and observability of conceptual rainfall-runoff model parameters: the percolation process examined, *Water Resources Research*, *19*(1), 269–276.

Gupta, V.K., and S. Sorooshian (1985), The relationship between data and the precision of parameter estimates of hydrological models, *Journal of Hydrology*, *81*, 57–77.

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

Gupta, H., M. Thiemann, M. Trosset, and S. Sorooshian (2003), Reply to comment by K.

Beven and P. Young on "Bayesian recursive parameter estimation for hydrologic models", *Water Resources Research*, *39*(5), 1117, doi:10.1029/2002WR001405.

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

Gupta, H.V., M.P. Clark, J.A. Vrugt, G. Abramowitz, and M. Ye (2012), Towards a comprehensive assessment of model structural adequacy, *Water Resour. Res.*, *48*, W08301, doi:10.1029/2011WR011044.

Haan, C.T. (1977), Statistical Methods in Hydrology, Iowa State University Press, Ames.

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

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

Hall, P., J.W. Kay, and D.M. Titterington (1990), Asymptotically optimal difference-based estimation of variance in nonparametric regression, *Biometrika*, *77*, 521–528.

Hankin B.G., R. Hardy, H. Kettle, and K.J. Beven (2001), Using CFD in a GLUE framework to model the flow and dispersion characteristics of a natural fluvial dead zone, *Earth Surface Processes and Landforms*, *26*, 667–687.

Hansen, T.M., K.S. Cordua, M.C. Looms, and K. Mosegaard (2012a), SIPPI: A Matlab toolbox for sampling the solution to inverse problems with complex prior information: Part 1-Methodology, *Computers & Geosciences*, *52*, 470–480.

Hansen, T.M., K.S. Cordua, M.C. Looms, and K. Mosegaard (2012b). SIPPI: A Matlab toolbox for sampling the solution to inverse problems with complex prior information: Part 2-Application to cross hole GPR tomography, *Computers & Geosciences*, *52*, 481–492.

Hansen, T.M., K.S. Cordua, and K. Mosegaard (2012c), Inverse problems with non-trivial priors: Efficient solution through sequential Gibbs sampling, *Computational Geosciences*, *16*(3), 593–611.

Hansson, K., and C. Lundin (2006), Equifinality and sensitivity in freezing and thawing simulations of laboratory and in situ data, *Cold Regions Science and Technology*, *44*, 20–37.

Hastings, H. (1970), Monte Carlo sampling methods using Markov chains and their applications, *Biometrika, 57*, 97–109.

Higuchi, T. (1997), Monte Carlo filter using the genetic algorithm operators, *Journal of Statistical Computation and Simulation*, *5*9, 1–23.

Hodgkins, G. A. and R. W. Dudley (2006), Changes in the timing of Winter-Spring streamflows in Eastern North America, 1913-2002, *Geophysical Research Letters*, *33*, L06402, doi: 10.1029/2005GL025593.

Holmes, M.G.R., A.R. Young, A. Gustard, and R. Grew (2002), A region of influence approach to predicting flow duration curves within ungauged catchments, *Hydrol. Earth Syst. Sci.*, *6*, 721-731, doi:10.5194/hess-6-721-2002.

Honarkhah, M., and J. Caers (2010), Stochastic simulation of patterns using distance-based pattern modeling, *Mathematical Geosciences*, *42*(5), 487–517.

Hong, T., and M. Sen (2009), A new MCMC algorithm for seismic waveform inversion and corresponding uncertainty analysis, *Geophysical Journal International*, *177*(1), 14–32.

Hu, L., and T. Chugunova (2008), Multiple-point geostatistics for modeling subsurface heterogeneity: A comprehensive review, *Water Resources Research*, *44*(11), W11413.

Huard, D., and A. Mailhot (2006), A Bayesian perspective on input uncertainty in model calibration: Application to hydrological model "abc", *Water Resources Research*, *42*, W07416, doi:10.1029/2005WR004661, 2006.

Hughes, D.A., and V. Smakhtin (1996), Daily flow time series patching or extension: a spatial interpolation approach based on flow duration curves, *Hydrological Sciences Journal*, *41*(6), 851-871, doi: 10.1080/02626669609491555.

Hurst, H.E. (1951), Long term storage capacities of reservoirs, *Transactions of the American Society of Civil Engineers*, *116*, 776–808.

Huynh-Thu, Q., and M. Ghanbari (2008), Scope of validity of PSNR in image/video quality assessment, *Electronics Letters*, *44*(13), 800–801.

Huysmans, M., and A. Dassargues (2009), Application of multiple-point geostatistics on modelling groundwater flow and transport in a cross-bedded aquifer (Belgium), *Hydrogeology Journal*, *17*(8), 1901–1911.

Iacobellis, V. (2008), Probabilistic model for the estimation of T year flow duration curves, *Water Resour. Res.*, *44*, W02413, doi:10.1029/2006WR005400.

Ibbitt, R.P. (1972), Effects of random data errors on the parameter values for a conceptual model, *Water Resources Research*, *8*(1), 70–78.

Ibbitt, R.P., and T. O'Donnell (1974), Designing conceptual catchment models for automatic fitting methods, *Mathematical Models in Hydrology Symposium, IAHS-AISH, Publication No. 2*, 461–475.

Irving, J., and K. Singha (2010), Stochastic inversion of tracer test and electrical geophysical data to estimate hydraulic conductivities, *Water Resources Research*, *46*(11), W11514.

Ishak, E.H., A. Rahman, S. Westra, A. Sharma, and G. Kuczera (2013), Evaluating the non-stationarity of Australian annual maximum flood, *Journal of Hydrology*, *494*, 134–145.

Jacquin A.P., and A.Y. Shamseldin (2007), Development of a possibilistic method for the evaluation of predictive uncertainty in rainfall-runoff modeling, *Water Resources Research*, *43*, W04425. doi:10.1029/2006WR005072.

Jafarpour, B. (2011), Wavelet reconstruction of geologic facies from nonlinear dynamic flow measurements. *IEEE Transactions on Geoscience and Remote Sensing*, *49*(5), 1520–1535.

Jafarpour, B., V. Goyal, D. McLaughlin, and W. Freeman (2009), Transform-domain sparsity regularization for inverse problems in geosciences, *Geophysics*, *74*(5), R69–R83.

Jafarpour, B., V. Goyal, D. McLaughlin, and W. Freeman (2010), Compressed history matching: Exploiting transform-domain sparsity for regularization of nonlinear dynamic data integration problems, *Mathematical Geosciences*, *42*(1), 1–27.

Jain, S., and U. Lall (2001), Floods in a changing climate: Does the past represent the future?, *Water Resources Research*, *37*(12), 3193-3205, doi:10.1029/2001WR000495.

Jehng-Jung, K., and S.F. Bau (1996), Risk Analysis for Flow Duration Curve Based Seasonal Discharge Management Programs, *Wat. Res.*, *30*(6), 1369–1376.

Jensen, J.B. (2003), Parameter and Uncertainty Estimation in Groundwater Modelling, PhD thesis, Department of Civil Engineering, Aalborg University, Series Paper No. 23.

Jimenez, R., L. Verde, H. Peiris, and A. Kosowsky (2004), Fast cosmological parameter estimation from microwave background temperature and polarization power spectra, *Physical Review D*, *70*(2), 23005.

Johnson, T., P. Routh, T. Clemo, W. Barrash, and W. Clement (2007), Incorporating geostatistical constraints in nonlinear inversion problems, *Water Resources Research*, *43*(10), W10422.

Johnston, P.R., and D. Pilgrim (1976), Parameter optimization for watershed models, *Water Resources Research*, *12*(3), 477–486.

Joshi, M. K., and A. C. Pandey (2011), Trend and spectral analysis of rainfall over India during 1901-2000, *Journal of Geophysical Research D: Atmospheres*, *116*(6), art. no. D06104.

Joyce, P. and P. Marjoram (2008), Approximately sufficient statistics and Bayesian computation, *Statistical Applications in Genetics and Molecular Biology*, *7* (1).

Kaipio, J.P., A. Seppänen, E. Somersalo, and H. Haario (2004), Posterior covariance related optimal current patterns in electrical impedance tomography, *Inverse Problems*, *20*, 919–936.

Karl, T.R., R.W. Knight (1998), Secular trends of precipitation amount, frequency, and intensity in the United States, *Bulletin of American Meteorological Society*, *79*(2), 231-241.

Karthikeyan, L., and D. Nagesh Kumar (2013), Predictability of nonstationary time series using wavelet and EMD based ARMA models, *Journal of Hydrology*, *502*, 103–119.

Katz, R.W., M.B. Parlange, P. Naveau (2002), Statistics of extremes in hydrology, *Advances in Water Resources*, *25*(812), 1287–1304.

Kavetski, D., S. W. Franks, and G. Kuczera (2002), Confronting input uncertainty in environmental modeling, In: Calibration of Watershed Models, Water Science Applications, vol. 6, Ed. Q. Duan et al., pp. 49–68, AGU, Washington, D. C.

Kavetski, D., G. Kuczera, and S.W. Franks (2003), Semi-distributed hydrological modeling: A "saturation path" perspective on TOPMODEL and VIC, *Water Resources Research*, *39* (9), 1246, doi:10.1029/2003WR002122.

Kavetski, D., G. Kuczera, and S.W. Franks (2006a), Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, *Water Resources Research*, *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 Resources Research*, *42*, W03408, doi:10.1029/2005WR004376.

Kavetski, D., G. Kuczera, and S.W. Franks (2006c), Calibration of conceptual hydrological models revisited: 1. Overcoming numerical artifacts, *Journal of Hydrology*, *320* (1-2), 173–186, doi:10.1016/j.jhydrol.2005.07.012.

Kavetski, D., and M.P. Clark (2010), Ancient numerical daemons of conceptual hydrological modeling: 2. Impact of time stepping schemes on model analysis and prediction, *Water Resources Research*, *46*, W10511, doi:10.1029/2009WR008896.

Kavetski, D., F. Fenicia, and M.P. Clark (2011), Impact of temporal data resolution on parameter inference and model identification in conceptual hydrological modeling: Insights from an experimental catchment, *Water Resour. Res.*, *47*, W05501, doi:10.1029/2010WR009525.

Keller, G.V. (1988), Rock and mineral properties. In Nabighian, M.N., editor, *Electromagnetic Methods in Applied Geophysics*, volume 1, pages 13–52. Society of Exploration Geophysicists.

Khaliq, M. N., T. B. M. J. Ouarda, P. Gachon, L. Sushama, A. St-Hilaire (2009), Identification of hydrological trends in the presence of serial and cross correlations: a review of selected methods and their application to annual flow regimes of Canadian rivers, *Journal of Hydrology*, *368*, 117-130.

Khu, S.T., and H. Madsen (2005), Multiobjective calibration with Pareto preference ordering: An application to rainfall-runoff model calibration, *Water Resources Research*, *41*, W03004, doi:10.1029/2004WR003041.

Kosugi, K. (1994), Three-parameter lognormal distribution model for soil water retention, *Water Resources Research*, *30*(4), 891–901.

Kosugi, K. (1996), Lognormal distribution model for unsaturated soil hydraulic properties, *Water Resources Research*, *32*(9), 2697–2703.

Kosugi, K., and J.W. Hopmans (1998), Scaling water retention curves for soils with lognormal pore-size distribution, *Soil Sci. Soc. Am. J.*, *62*, 1496–1505.

Koutsoyiannis, D. (2006), Nonstationarity versus scaling in hydrology, *Journal of Hydrology*, *324*, 239-254.

Koutsoyiannis, D. (2011), Hurst-Kolmogorov dynamics and uncertainty, *JAWRA Journal of the American Water Resources Association*, *47*(3), 481–495.

Koutsoyiannis, D. (2013), Hydrology and change, *Hydrological Sciences Journal*, *58*(6), 1177–1197, doi: 10.1080/02626667.2013.804626

Kuczera, G. (1982), On the relationship between the reliability of parameter estimates and hydrologic time series used in calibration, *Water Resources Research*, *18*(1), 146–154.

Kuczera, G. (1983a), Improved parameter inference in catchment models, 1. Evaluating parameter uncertainty, *Water Resources Research*, *19*(5), 1151–1162, doi:10.1029/WR019i005p01151.

Kuczera, G. (1983b), Improved parameter inference in catchment models: 2. Combining different kinds of hydrologic data and testing their compatibility, *Water Resources Research*, *19*(5), 1163–1172.

Kuczera, G., and E. Parent (1998), Monte Carlo assessment of parameter uncertainty in conceptual catchment models: The Metropolis algorithm, *Journal of Hydrology*, *211*(1-4), 69–85.

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, *Journal of Hydrology*, *331*, 161–177, doi:10.1016/j.jhydrol.2006.05.010.

Kuczera, G., D. Kavetski, B. Renard, and M. Thyer (2010), A limited-memory acceleration strategy for MCMC sampling in hierarchical Bayesian calibration of hydrological models, *Water Resources Research*, *46*, W07602, doi:10.1029/2009WR008985.

Kundzewicz, Z. W., A. J. Robson (2004), Change detection in hydrological records: a review of the methodology, *Hydrological Sciences*, *49*, 7-19.

Kundzewicz, Z. W., D. Graczyk , T. Maurer , I. Pińskwar , M. Radziejewski , C. Svensson, and M. Szwed (2005), Trend detection in river flow series: 1. Annual maximum flow, *Hydrological Sciences Journal*, *50*(5), 797–810, doi: 10.1623/hysj.2005.50.5.797

Kundzewicz, Z. W. (2011), Nonstationarity in water resources central European perspective, *JAWRA Journal of the American Water Resources Association*, *47*, 550-562.

Kunkle, G.R. (1962), The Baseflow-Duration Curve, a Technique for the study of Groundwater Discharge from a Drainage Basin, *Journal of Geophysical Research*, *67*(4), 1543–1554.

Lagarias, J.C., J.A. Reeds, M.H. Wright, and P.E. Wright (1998), Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions, *SIAM Journal of Optimization*, *9*(1), 112–147.

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

Laloy, E., J.A. Vrugt, and G. Schoups (2012), Bayesian model selection using the marginal likelihood, *Water Resources Research,*, In Review.

Lamb R, K. Beven, and S. Myrabø (1998), Use of spatially distributed water table observations to constrain uncertainty in a rainfall-runoff model, *Advances in Water Resources*, *22*(4), 305–317.

Lane, P.N.J., A.E. Best, K. Hickel, and L. Zhang (2005), The response of flow duration curves to afforestation, *Journal of Hydrology*, *310*, 253–265.

Lange, K., K. Cordua, J. Frydendall, T. Hansen, and K. Mosegaard (2012), A frequency matching method for generation of a priori sample models from training images, *Mathematical Geosciences*, *44*, 783–803.

Leboutillier, D.W., and P.R. Waylen (1993), A Stochastic Model of Flow Duration Curves, *Water Resources Research*, *29*(10), 3535–3541.

Leclerc, M., and T. B. M. J. Ouarda (2007), Non-stationary regional flood frequency analysis at ungauged sites, *Journal of Hydrology*, *343*, 254-265.

Leopold, L. B. (1968), Hydrology for urban planning – A guidebook on the hydrologic effects of urban land use, *U.S. Geological Survay Circular*, 554.

Lettenmaier, D. P., E. F. Wood, J. R. Wallis (1994), Hydroclimatological trends in the continental United States, 1948–1988, *Journal of Climate*, *7*, 586-607

Lewis, S.M., and A.E. Raftery (1997), Estimating Bayes' factors via posterior simulation with the Laplace-Metropolis estimator, *Journal of the American Statistical Association*, *92*, 648–655.

Li, M., Q. Shao, L. Zhang, and F.H.S. Chiew (2010), A new regionalization approach and its application to predict flow duration curve in ungauged basins, *J. Hydrol.*, *389*(12), 137–145.

Lima, C. H. R., and U. Lall (2010), Spatial scaling in a changing climate: A hierarchical

bayesian model for non-stationary multi-site annual maximum and monthly streamflow, *Journal of Hydrology*, *383*(3-4), 307–318.

Linde, N., A. Binley, A. Tryggvason, L.B. Pedersen, and A. Revil (2006), Improved hydrogeophysical characterization using joint inversion of cross-hole electrical resistance and ground-penetrating radar traveltime data, *Water Resources Research*, *42*(12), W04410.

Linde, N., and J.A. Vrugt (2013), Distributed soil moisture from crosshole ground-penetrating radar travel times using stochastic inversion, *Vadose Zone Journal*, *12*(1).

Lins, H. F. (1985), Streamflow variability in the United States: 1931-78. *Journal of Climate Applied Meteology*, *24*, 463-471.

Lins, H. F. and T. A. Cohn (2011), Stationarity: wanted dead or alive?, *JAWRA Journal of the American Water Resources Association*, *47*(3), 475–480.

Lins, H. F., and J. R., Slack (1999), Streamflow trends in the United States, *Geophysical Research Letters*, *26*(2), 227–230.

Lins, H. F., and J. R. Slack (2005), Seasonal and regional characteristics of U.S. streamflow trends in the United States from 1940–1999. *Physical Geography*, *26*, 489–501.

Lochbühler, T., S.J. Breen, R.L. Detwiler, J.A. Vrugt, and N. Linde (2014a), Probabilistic electrical resistivity tomography of a CO_2 sequestration analog, *Journal of Applied Geophysics*, *107*, 80–92.

Lochbühler, T., G. Pirot, J. Straubhaar, and N. Linde (2014b), Conditioning of multiple-point statistics facies simulations to tomographic images, *Mathematical Geosciences*, *46*, 625–645.

Lochbühler, T., J.A. Vrugt, M. Sadegh, and N. Linde (2014c), Summary Statistics from Training Images as Prior Information in Probabilistic Inversion, *Geophysical Research Letters*, Under Review.

Male, J.W., and H. Ogawa (1984), Tradeoffs in water quality management, *Water Resources Planning and Management*, *110*(4), 434–444.

Malinverno, A. (2002), Parsimonious Bayesian Markov chain Monte Carlo inversion in a nonlinear geophysical problem, *Geophysical Journal International*, *151*, 675–688.

Malinverno, A., and V.A. Briggs (2004), Expanded uncertainty quantification in inverse problems: Hierarchical Bayes and empirical Bayes, *Geophysics*, *69*(4), 1005–1016.

Mallory, S.J.L., and R.S. McKenzie (1993), Water resources modeling of flow diversions. *Proceedings of the Sixth South African National Hydrology Symposium, Pietermarietzburg, South Africa*, *1*, 429–436.

Mantovan, P., and E. Todini (2006), Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology, *Journal of Hydrology*, *330*, 368–381, doi:10.1016/j.jhydrol.2006.04.046.

Mariethoz, G., P. Renard, and J. Caers (2010a), Bayesian inverse problem and optimization with iterative spatial resampling, *Water Resources Research*, *46*(11), W11530.

Mariethoz, G., P. Renard, and J. Straubhaar (2010b), The direct sampling method to perform multiple-point geostatistical simulations, *Water Resources Research*, *46*(11), W11536.

Marjoram, P., J. Molitor, V. Plagnol, and S. Tavaré (2003), Markov chain Monte Carlo without likelihoods, *Proceedings of the National Academy of Sciences of the United States of America*, *100* (26), 15324–15328.

Marquardt, D.W. (1963), An algorithm for least-squares estimation of nonlinear parameters, *Journal of the Society for Industrial & Applied Mathematics*, *11*(2), 431–441.

Marshall, L., D. Nott, and A. Sharma (2004), A comparative study of Markov chain Monte

Carlo methods for conceptual rainfall-runoff modeling, *Water Resources Research*, *40*(2), 1–11.

Marshall, L.A., A. Sharma, and D.J. Nott (2007), Towards dynamic catchment modelling: A Bayesian hierarchical mixtures of experts framework, *Hydrological Processes*, *21*, 847–861.

Masih, I., S. Uhlenbrook, S. Maskey, M.D. Ahmad (2010), Regionalization of a conceptual rainfallrunoff model based on similarity of the flow duration curve: A case study from the semi-arid Karkheh basin, Iran, *Journal of Hydrology*, *391*, 188–201.

Maurer, H., K. Holliger, and D. Boerner (1998), Stochastic regularization: Smoothness or similarity?, *Geophysical Research Letters*, *25*(15), 2889–2892.

Mazi, K., A.D. Koussis, P.J. Restrepo, and D. Koutsoyiannis (2004), A groundwater-based, objective-heuristic parameter optimization method for a precipitation-runoff model and its application to a semi-arid basin, *Journal of Hydrology*, *290*, 243–258.

McCabe, G. J., and D. M. Wolock (2002), A step increase in streamflow in the coterminous United States, *Geophysical Research Letters*, *29*(24), 2185, doi: 10.1029/2002GL015999.

McClymont, A., A. Green, R. Streich, H. Horstmeyer, J. Tronicke, D. Nobes, J. Pettinga, J. Campbell, and R. Langridge (2008), Visualization of active faults using geometric attributes of 3D GPR data: An example from the Alpine Fault Zone, New Zealand. *Geophysics*, *73*(2), B11–B23.

McMichael, C.E., A.S. Hope, and H.A. Loaiciga (2006), Distributed hydrological modeling in California semi-arid shrublands: MIKE SHE model calibration and uncertainty estimation, *Journal of Hydrology*, *317*, 307–324.

Meerschman, E., G. Pirot, G. Mariethoz, J. Straubhaar, M. Van Meirvenne, and P. Renard (2013), A practical guide to performing multiple-point statistical simulations with the direct sampling algorithm. *Computers & Geosciences*, *52*, 307–324.

Gelman, A., and X. Meng (1998), Simulating normalizing constants: From importance sampling to bridge sampling to path sampling, *Statistical Science*, *13* (2), 163–185.

Mertens, J., H. Madsen, L. Feyen, D. Jacques, J. and Feyen (2004), Including prior information in the estimation of effective soil parameters in unsaturated zone modelling, *Journal of Hydrology*, *294*(4), 251–269.

Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller (1953), Equation of state calculations by fast computing machines, *Journal of Chemical Physics*, *21*, 1087–1092.

Miller, J. R., and G. L. Russell (1992), The impact of global warming on river runoff, *Journal of Geophysical Research*, *97*(D3), 2757–2764.

Milly, P. C. D., J. Betancourt, M. Falkenmark, R. M. Hirsch, Z. W. Kundzewicz, D. P. Lettenmaier, and R. J. Stouffer (2008), Stationarity is dead: Whiter water management?, *Science*, *319*, 573-574.

Mimikou, M., and S. Kaemaki (1985), Regionalization of Flow Duration Characteristics, *Journal of Hydrology*, *82*, 77–91.

Mitchell, W.D. (1957), Flow duration curves of Illinois streams, Illinois Department of Public Works and Buildings, Division of Waterways.

Mohamoud, Y. M. (2008), Prediction of daily flow duration curve and streamflow for ungauged catchments using regional flow duration curves, *Hydrolog. Sci. J.*, *53*, 706–724, doi:10.1623/hysj.53.4.706.

Mohanty, B.P., R.S. Bowman, J.M.H. Hendrickx, and M.T. van Genuchten (1997), New piecewise-continuous hydraulic functions for modeling preferential flow in an intermittent-flood-irrigated field, *Water Resources Research*, *33*(9), 2049–1063.

Montanari, A., and A. Brath (2004), A stocastic approach for assessing the uncertainty of rainfall-runoff simulations, *Water Resources Research*, *40*, W01106, doi:10.1029/2003WR002540.

Montanari, A. (2005), Large sample behaviors of the generalized likelihood uncertainty estimation(GLUE) in assessing the uncertainty of rainfall-runoff simulations, *Water Resources Research*, *41*, W08406, doi:10.1029/2004WR003826.

Montanari A., and G. Grossi (2008), Estimating the uncertainty of hydrological forecasts: A statistical approach, *Water Resources Research*, *44*, W00B08, doi:10.1029/2008WR006897.

Moradkani, H., K. Hsu, H. Gupta, and S. Sorooshian (2005a), Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter, *Water Resources Research*, *41*, W05012, doi:10.1029/2004WR003604.

Moradkhani, H., S. Sorooshian, H. Gupta, and P.R. Houser (2005b), Dual stateparameter estimation of hydrological models using ensemble Kalman filter, *Advances in Water Resources*, *28*, 135–147.

Mosegaard, K., and A. Tarantola (1995), Monte Carlo sampling of solutions to inverse problems, *Journal of Geophysical Research*, *100*(B7), 12431–12447.

Morin, E., K. P. Georgakakos, U. Shamir, R. Garti, and Y. Enzel (2002), Objective, observations-based, automatic estimation of the catchment response timescale, Water Resour. Res., 38(10), 1212, doi:10.1029/2001WR000808.

Mroczkowski, M., G. P. Raper, and G. Kuczera (1997), The quest for more powerful validation of conceptual catchment models, *Water Resources Research*, *33*(10), 2325–2335.

Muleta, M.K., and J.W. Nicklow (2005), Sensitivity and uncertainty analysis coupled with

automatic calibration for a distributed watershed model, *Journal of Hydrology*, *306*, 127–145.

Nash, J.E. (1960), A unit hydrograph study with particular reference to British catchments, *Proceedings - Institution of Civil Engineers*, *17*, 249–282.

Nash, J.E., and J.V. Sutcliffe (1970), River flow forecasting through conceptual models part I - A discussion of principles, *Journal of Hydrology*, *10*(3), 282–290.

Nasri, B., S. El-Adlouni, and T. B. M. J. Ouarda (2013), Bayesian estimation for GEV-B-Spline model, *Open Journal of Statistics*, *3*(2), 118–128.

Nelder, J.A., and R. Mead (1965), A simplex method for function minimization, *Computer Journal*, *7*, 308–313.

Ni, C.F., C. P. Lin, S. G. Li, and C. J. Liu (2011), Efficient approximate spectral method to delinate stochastic well capture zones in nonstationary groundwater flow systems, *Journal of Hydrology*, *407*(1-4), 184–195.

Niadas, I.A. (2005), Regional flow duration curve estimation in small ungauged catchments using instantaneous flow measurements and a censored data approach, *Journal of Hydrology*, *314*, 48–66.

Nissinen, A., L.M. Heikkinen, V. Kolehmainen, and J.P. Kaipio (2009), Compensation of errors due to discretization, domain truncation and unknown contact impedances in electrical impedance tomography, *Measurement Science and Technology*, *20*, 105504, doi:10.1088/0957-0233/20/10/105504.

North, M. (1980), Time-dependent stochastic model of floods, *Journal of the Hydraulics Division*, *106*(5), 649–665.

Nott, D.J., L. Marshall, and J. Brown (2012), Generalized likelihood uncertainty estimation (GLUE) and approximate Bayesian computation: What's the connection?, *Water Resources Research*, *48*(12), doi:10.1029/2011WR011128

Novotny, E. V., H. G. Stefan (2007), Stream flow in Minnesota: indicator of climate change, *Journal of Hydrology*, *334*, 319-333.

Olden, J. D., and N. L. Poff (2003), Redundancy and the choice of hydrologic indices for characterizing streamflow regimes, *River Research and Applications*, *19*, 101–121.

Ouarda, T. B. M. J., and S. El-Adlouni (2011), Bayesian nonstationary frequency analysis of hydrological variables. *JAWRA Journal of the American Water Resources Association*, *47*(3), 496–505.

Parey S., F. Malek, C. Laurent, D. Dacunha-Castelle (2007), Trends and climate evolution: Statistical approach for very high temperatures in France, *Climatic Change*, *81*(3-4), 331–352.

Pauwels, V.R.N. (2008), A multi-start weight-adaptive recursive parameter estimation method, *Water Resources Research*, *44*(4), W044156, doi:10.1029/2007WR005866.

Peterson, J.E.J. (2001), Pre-inversion processing and analysis of tomographic radar data, *Journal of Environmental & Engineering Geophysics*, *6*(1), 1–18.

Petrow, T., B. Merz (2009), Trends in flood magnitude, frequency and seasonality in Germany in the period 1951-2002, *Journal of Hydrology*, *371*, 129-141.

Pfannerstill, M., B. Guse, and N. Fohrer (2014), Smart low flow signature metrics for an improved overall performance evaluation of hydrological models, *Journal of Hydrology*, *510*, 447–458.

Pitman, W. V. (1978), Trends in streamflow due to upstream land-use changes, *Journal of Hydrology*, *39*, 227–237.

Pitman, W.V. (1993), Simulation of run-of-river schemes using monthly data, *Proceedings of the Sixth South African Hydrology Symposium, Pietermarietzburg, South Africa*, *1*, 445–452.

Podvin, P., and I. Lecomte (1991), Finite difference computation of traveltimes in very contrasted velocity models: A massively parallel approach and its associated tools, *Geophysical Journal International*, *105*(1), 271–284.

Potter, K. W. (1991), Hydrological impacts of changing land management practices in a moderate-sized agricultural catchment, *Water Resources Research*, *27*(5), 845-855.

Price, K.V., Storn, R.M., and J.A. Lampinen (2005), *Differential evolution, A practical approach to global optimization*, Springer, Berlin.

Pride, S. (1994), Governing equations for the coupled electromagnetics and acoustics of porous media, *Physical Review B*, *50*(21), 15678.

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, *Molecular Biology and Evolution*, *16*(12), 1791–1798.

Quimpo, R.G., A.A. Alejandrino, and T.A. McNally (1983) Regionalised flow duration curves for Philippines, *Journal of Water Resources Planning and Management*, *109*(4), 320330.

Ramachandra Rao, A., and G. H. Yu (1986), Detection of nonstationarity in hydrologic time series, *Management Science*, *32*(9), 1206–1217.

Ratmann, O., C. Andrieu, C. Wiuf, and S. Richardson (2009), Model criticism based on likelihood-free inference, with an application to protein network evolution, *Proceedings of the National Academy of Sciences of the United States of Ameria*, *106*, 1–6.

Refsgaard, J.C. and J. Knudsen (1996), Operational validation and intercomparison of different types of hydrological models, *Water Resour. Res.*, *32*, 21892202.

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

Reichle, R.H. (2008), Data assimilation methods in the earth sciences, *Advances in Water Resources*, *31*, 1411–1418.

Renard, B., M., Lang, and P. Bois (2006), Statistical analysis of extreme events in a non-stationary context via a Bayesian framework: case study with peak-over-threshold data, *Stochastic Environmental Research and Risk Assessmement*, *21*, 97-112.

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

Renard, P., J. Straubhaar, J. Caers, and G. Mariethoz (2011), Conditioning facies simulations with connectivity data, *Mathematical Geosciences*, *43*, 879–903.

Restrepo-Posada, J.P. (1982), Automatic parameter estimation of a large conceptual rainfall-runoff model: a maximum likelihood approach, Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, pp. 330.

Rice, J. (1984), Bandwidth choice for nonparametric kernel regression, *Annals of Statistics, 12*, 1215–1230.

Rings, J., J.A. Vrugt, G. Schoups, J.A. Huisman, and H. Vereecken (2012), Bayesian model averaging using particle filtering and Gaussian mixture modeling: Theory, concepts, and simulation experiments, *Water Resources Research*, doi:10.1029/2011WR011607.

Romanowicz R.J., K.J. Beven, J. and Tawn (1996), Bayesian calibration of flood inundation

models, M.G. Anderson and D.E. Walling(eds.), Floodplain Processes, Wiley, Chichester, pp. 333–360.

Rosas-Carbajal, M., N. Linde, T. Kalscheuer, and J.A. Vrugt (2014), Two-dimensional probabilistic inversion of plane-wave electromagnetic data: Methodology, model constraints and joint inversion with electrical resistivity data, *Geophysical Journal International*, *196*(3), 1508–1524.

Rougé, C., Y. Gea, X. Cai (2013), Detecting gradual and abrupt changes in hydrological records, *Advances in Water Resources*, *53*, 33-44.

Rudin, L.I., S. Osher, and E. Fatemi (1992), Nonlinear total variation based noise removal algorithms, *Physica D: Nonlinear Phenomena*, *60*(1), 259–268.

Ruggeri, P., J. Irving, and K. Holliger (2013), Systematic evaluation of geostatistical resampling within MCMC for generating Bayesian posterior realizations of hydrogeophysical problems. In *AGU Fall Meeting Abstracts*, Abstract ID: H51G-1275.

Sadegh, M., and J.A. Vrugt (2013), Bridging the gap between GLUE and formal statistical approaches: approximate Bayesian computation, *Hydrology and Earth System Sciences*, *17*, 4831-4850, doi:10.5194/hess-17-4831-2013.

Sadegh, M., and J.A. Vrugt (2014), Approximate Bayesian Computation using Markov Chain Monte Carlo simulation: DREAM$_{(\mathrm{ABC})}$, *Water Resour. Res.*, *50*, 67676787, doi:10.1002/2014WR015386.

Sadegh, M., J.A. Vrugt, C. Xu, and E. Volpi (2015a), Diagnosing Nonstationary Watershed Behavior from Process-Based Model Evaluation using Approximate Bayesian Computation, *Water Resources Research*, under review.

Sadegh, M., J.A. Vrugt, H.V. Gupta (2015b), The soil water characteristic as new class of

closed-form parametric expressions for the flow duration curve, *Water Resources Research*, under review.

Salamon, P., and L. Feyen (2009), Assessing parameter, precipitation, and predictive uncertainty in a distibuted hydrological model using sequential data assimilatiom with the particle filter, *Journal of Hydrology*, *376*, 428–442.

Salas, J. D. (1993), Analysis and modeling of hydrologic time series. In: Maidment, D. (Ed.), *Handbook of Hydrology*, McGraw-Hill, New York, pp 19.1–19.72 (Chapter 19).

Salas, J. D., and J. Obeysekera (2014), Revisiting the concepts of return period and risk for nonstationary hydrologic extreme events. *Journal of Hydrologic Engineering*, *19*(3), 554–568.

Sambridge, M., and K. Mosegaard (2002), Monte Carlo methods in geophysical inverse problems, *Reviews of Geophysics*, *40*(3), 1009.

Sauer, V. B., W. O. Thomas, V. A. Stricker, and K. V. Wilson (1983), Flood characteristics of urban watersheds in the United States, *U.S. Geological Survey Water Supply Paper*, 2207.

Sauquet, E., and C. Catalogne (2011), Comparison of catchment grouping methods for flow duration curve estimation at ungauged sites in France, *Hydrol. Earth Syst. Sci.*, *15*, 2421-2435.

Savenije, H.H.G. (1996), The runoff coefficient as the key to moisture recycling, *Journal of Hydrology*, *176*, 219-225.

Savicz, K., T., Wagener, M., Sivapalan, P. A., Troch, and G. Carrillo (2011), Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA, *Hydrology and Earth System Sciences*, *15*, 2895–2911.

Schaake, J., S. Cong, Q. Duan (2006), U.S. MOPEX data set, *IAHS Publication Series*, available at `http://goo.gl/wthGvS`.

Schaefli, B., and H.V. Gupta (2007), Do Nash values have value?, *Hydrol. Processes*, *21*, 2075–2080.

Scharnagl, B., J.A. Vrugt, H. Vereecken, and M. Herbst (2011), Bayesian inverse modeling of soil water dynamics at the field scale: using prior information about the soil hydraulic properties, *Hydrology and Earth System Sciences*, *15*, 3043-3059, doi:10.5194/hess-15-3043-2011.

Scholer, M., J. Irving, M. Looms, L. Nielsen, and K. Holliger (2012), Bayesian Markov-Chain-Monte-Carlo inversion of time-lapse crosshole GPR data to characterize the vadose zone at the Arrenaes site, Denmark, *Vadose Zone Journal*, *11*(4).

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

Schoups, G., J.A. Vrugt, F. Fenicia, and N.C. van de Giesen (2010b), Corruption of accuracy and efficiency of Markov Chain Monte Carlo simulation by inaccurate numerical implementation of conceptual hydrologic models, Water Resources Research, 46, W10530, doi:10.1029/2009WR008648.

Schoups, G., and J.A. Vrugt (2011), Bayesian inference of the marginal density of hydrologic models: Path-sampling using DREAM, *Water Resources Research*, In Review.

Searcy, J.K. (1959), Flow-duration curves, *Geological water supply paper: Methods and practices of the Geological Survey*, pp. 1–33.

Seifert, B., T. Gasser, and A. Wolf (1993), Nonparametric estimation of residual variance revisited, *Biometrika*, *80*, 373–383.

Seibert, J. (2000), Multi-criteria calibration of a conceptual rainfall-runoff model using a genetic algorithm, *Hydrology and Earth System Sciences*, *4*(2), 215–224.

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, *Hydrological Processes*, *19*, 2187–2207.

Singh, K. P. (1971), Model flow duration and streamflow variability, *Water Resour. Res.*, *7*(4), 1031–1036.

Singh, R.D., S.K. Mishra, and H. Chowdhary (2001), Regional flow-duration models for large number of ungauged Himalayan catchments for planning microhydro projects, *J. Hydrol. Eng.*, *6*(4), 310–316.

Sisson, S.A., Y. Fan, and M.M. Tanaka (2007), Sequential Monte Carlo without likelihoods, *Proceedings of the National Academy of Sciences of the United States of America*, *104*(6), 1760–1765.

Sivapalan, M., K. Takeuchi, S.W. Franks, V.K. Gupta, H. Karambiri, V. Lakshmi, X. Liang, J.J. McDonnell, E.M. Mendiondo, P. E. O'Connell, T. Oki, J.W. Pomeroy, D. Schertzer, S. Uhlenbrook, and E. Zehe (2003), IAHS decade on predictions in ungauged basins (PUB), 20032012: Shaping an exciting future for the hydrological sciences, *Hydrolog. Sci. J.*, *48*, 857–880, doi:10.1623/hysj.48.6.857.51421.

Slater, A.G., and M.P. Clark (2006), Snow data assimilation via an ensemble Kalman filter, *Journal of Hydrometeorology*, *7*(3), 478–493.

Smakhtin, V. U. (2001), Low flow hydrology: a review, *J. Hydrol.*, *240*, 147–186.

Smith, J. A., M. L. Baeck, J. E. Morrison, P. Sturdevant-Rees, D. F. Turner-Gillespie, and P. D. Bates (2002), The regional hydrology of extreme floods in an urbanizing drainage, *Journal of Hydrometeorology*, *3*(3), 267-282.

Smith, P., K.J. Beven, and J.A. Tawn (2008), Detection of structural inadequacy in process-based hydrological models: a particle-filtering approach, *Water Resources Research*, *44*(1), W01410, doi:10.1029/2006WR005205.

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 Resources Research*, *46*(12), W12551.

Solomatine, D.P., and D.L. Shrestha (2009), A novel method to estimate model uncertainty using machine learning techniques, *Water Resources Research*, *45*, W00B11, doi:10.1029/2008WR006839.

Son, K. and M. Sivapalan (2007), Improving model structure and reducing parameter uncertainty in conceptual water balance models through the use of auxiliary data, *Water Resour. Res.*, *43*, W01415, doi:01410.01029/02006wr005032.

Sorooshian, S., and J.A. Dracup (1980), Stochastic parameter estimation procedures for hydrologic rainfall-runoff models: Correlated and heteroscedastic error cases, *Water Resources Research*, *16*(2), 430–442.

Sorooshian, S., V.K. Gupta, and J.L. Fulton (1983a), Evaluation of maximum likelihood parameter estimation techniques for conceptual rainfall-runoff models: influence of calibration data variability and length on model credibility, *Water Resources Research*, *19*(1), 251–259.

Sorooshian, S. and V.K. Gupta (1983b), Automatic calibration of conceptual rainfall-runoff models: the question of parameter observability and uniqueness, *Water Resources Research*, *19*(1), 260–268.

Stedinger, J.R., R.M. Vogel, S.U. Lee, and R. Batchelder (2008), Appraisal of Generalized Likelihood Uncertainty Estimation(GLUE) Methodology, *Water Resources Research*, *44*, WOOB06, doi:10.1029/2008WR006822.

Stedinger, J. R., and V. W. Griffis (2011), Getting from here to where? Flood frequency analysis and climate, *JAWRA Journal of the American Water Resources Association*, *47*, 506-513.

Stenerud, V.R., V. Kippe, K. Lie, A. Datta-Gupta (2008), Adaptive multiscale streamline simulation and inversion for high-resolution geomodels, *Society of Petroleum Engineers Journal*, *13*(1), 99–111.

Storn, R., and K. Price (1997), Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces, *Journal of Global Optimization*, *11*, 341–359.

Strebelle, S. (2002), Conditional simulation of complex geological structures using multiple-point statistics. *Mathematical Geology*, *34*(1), 1–21.

Strupczewski, W. G., V. P. Singh, and H. T. Mitosek (2001), Non-stationary approach to at-site flood frequency modelling. III. Flood analysis of Polish rivers, *Journal of Hydrology*, *248*(1-4), 152–167.

Sunnåker M, A.G. Busetto, E. Numminen J. Corander M. Foll, and C. Dessimoz (2013), Approximate Bayesian Computation, *Plos Computational Biology*, *9*(1), e1002803, 1–10, doi:10.1371/journal.pcbi.1002803.

Svensson, C., W. Z. Kundzewicz, and T. Maurer (2005), Trend detection in river flow series: 2. Flood and low-flow index series, *Hydrological Sciences Journal*, *50*(5), 811–824.

Tarantola, A. (2005), *Inverse Problem Theory and Methods for Model Parameter Estimation*, Society for Industrial Mathematics.

ter Braak, C.J.F., and J.A. Vrugt (2008), Differential evolution Markov chain with snooker updater and fewer chains, *Statistics and Computing*, *18*(4), 435–446, doi:10.1007/s11222-008-9104-9.

Tharme, R.E. (2003), A global perspective on environmental flow assessment: emerging trends in the development and application of environmental flow methodologies for rivers, *River Res. Appl.*, *19*, 397-441, doi:10.1002/rra.736.

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 Resources Research*, *45*, W00B14, doi:10.1029/2008WR006825.

Tadesse, A., and E.N., Anagnostou (2005), A statistical approach to ground radar-rainfall estimation, *Journal of Atmospheric and Ocean Technology*, *22*(11), 1055–1071.

Tolson, B.A., and C.A. Shoemaker (2007), Dynamically dimensioned search algorithm for computationally efficient watershed model calibration, *Water Resources Research*, *43*, W01413, doi:10.1029/2005WR004723.

Toni, T., D. Welch, N. Strelkowa, A. Ipsen, and M.P. Stumpf (2009), Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems, *Journal of the Royal Society Interface*, *6*, 187–202.

Troutman, B.M. (1985a), Errors and parameter estimation in precipitation-runoff modeling: I. Theory, *Water Resources Research*, *21*(8), 1195–1213.

Troutman, B.M. (1985b), Errors and parameter estimation in precipitation-runoff modeling: II. Case study, *Water Resources Research*, *21*(8), 1214–1222.

Turner, B.M, and P.B. Sederberg (2012), Approximate Bayesian computation with differential evolution, *Journal of Mathematical Psychology*, *56*(5), 375–385, doi:10.1016/j.jmp.2012.06.004.

Turner, B.M, and T. van Zandt (2012), A tutorial on approximate Bayesian computation, *Journal of Mathematical Psychology*, *56*, 69–85.

van Genuchten, M.T. (1980), A closed-form equation for predicting the hydraulic conductivity of unsaturated soils, *Soil Science Society of America Journal*, *44*(5), 892–898.

Vaze, J., D. A. Post, F. H. S. Chiew, J. -M. Perraud, N. R. Viney, and J. Teng (2010), Climate non-stationarity – Validity of calibrated rainfallrunoff models for use in climate change studies, *Journal of Hydrology*, *394*, 447-457.

Villarini, G., F. Serinaldi, J. A. Smith, and W. F. Krajewski (2009), On the stationarity of annual flood peaks in the continental United States during the 20th century, *Water Resources Research*, *45*, W08417.

Villarini, G., and J. A. Smith (2010), Flood peak distributions for the eastern United States, *Water Resources Research*, *46*, W06504, doi:10.1029/2009WR008395.

Vogel, R.M., and N.M. Fennessey (1994), Flow-duration curves I: New interpretation and confidence intervals, *Journal of Water Resources Planning and Management*, *120*(4), 485–504.

Vogel, R.M., and N.M. Fennessey (1995), Flow-duration curves II: A Review of applications in water resources planning, *Water Resour. Bull.*, *31*(6), 1029–1039.

Vogel, R.M., C. Yaindl, and M. Walter (2011), Nonstationarity: Flood magnification and recurrence reduction factors in the United States, *JAWRA Journal of the American Water Resources Association*, *47*(3), 464-474.

Vrugt, J.A., W. Bouten, H.V. Gupta, and S. Sorooshian (2002), Toward improved identifiability of hydrologic model parameters: The information content of experimental data, *Water Resources Research*, *38*(12), 1312, doi:10.1029/2001WR001118.

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 Resources Research, 41(1)*, W01017, doi:10.1029/2004WR003059.

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

Vrugt, J.A., and B.A. Robinson (2007b), Improved evolutionary optimization from genetically adaptive multimethod search, *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 708–711.

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 Resources Research*, *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 Environmental Research and Risk Assessment, 23(7)*, 1011–1026, doi:10.1007/s00477-008-0274-y.

Vrugt, J.A., C.J.F. ter Braak, H.V. Gupta, and B.A. Robinson (2008c), Response to comment by Keith Beven on "Equifinality of formal.(DREAM) and informal.(GLUE) Bayesian approaches in hydrologic modeling?", *Stochastic Environmental Research and Risk Assessment*, doi:10.1007/s00477-008-0284-9.

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, *International Journal of Nonlinear Sciences and Numerical Simulation, 10(3)*, 273–290.

Vrugt, J.A., B.A. Robinson, and J.M. Hyman (2009b), Self-adaptive multimethod search for global optimization in real-parameter spaces , *IEEE Transactions on Evolutionary Computation*, *13*(2), 243–259.

Vrugt, J.A., and C.J.F. ter Braak, (2011), DREAM$_{(D)}$: an adaptive Markov Chain Monte Carlo simulation algorithm to solve discrete, noncontinuous, and combinatorial posterior parameter estimation problems, *Hydrology and Earth System Sciences*, *15*, 3701–3713, doi:10.5194/hess-15-3701-2011.

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, *Advances in Water Resources, xx*, doi:10.1016/j.advwatres.2012.04.002.

Vrugt, J.A., and M. Sadegh (2013), Toward diagnostic model calibration and evaluation: Approximate Bayesian computation, *Water Resources Research*, *49*, doi:10.1002/wrcr.20354.

Waage, M. D., and L. Kaatz (2011), Nonstationary water planning: An overview of several promising planning methods. *JAWRA Journal of the American Water Resources Association*, *47*, 535-540.

Wagener, T., N. McIntyre, M.J. Lees, H.S. Wheater, and H.V. Gupta (2003), Towards reduced uncertainty in conceptual rainfall-runoff modelling: Dynamic identifiability analysis, *Hydrological Processes*, *17*, 455–476.

Wagener, T., and H.S. Wheater (2006), Parameter estimation and regionalization for continuous rainfall-runoff models including uncertainty, *Journal of Hydrology*, *320*, 132–154.

Wang, Q. (1991), The genetic algorithm and its application for calibrating conceptual rainfall-runoff models, *Water Resources Research*, *27*(9), 2467–2471.

Wang, X., X. He, J.R. Williams, R.C. Izaurralde, and J.D. Atwood (2005), Sensitivity and

uncertainty analyses of crop yields and soil organic carbon simulated with EPIC, *Transactions of the American Society of Agricultural Engineers*, *48*(3), 1041–1054.

Warnick, C.C. (1984), Hydropower engineering, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 59–73.

Wegmann, D., C. Leuenberger, and L. Excoffier (2009), Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood, *Genetics*, *182*(4), 1207–1218.

Welstead, S.T. (1999), *Fractal and Wavelet Image Compression Techniques*, SPIE Optical Engineering Press, Bellingham, WA.

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*, 22052227, doi:10.5194/hess-15-2205-2011.

Westmacott, J. R., and D. H. Burn (1997), Climate change effects on the hydrologic regime within the Churchill-Nelson River basin, *Journal of Hydrology*, *202*(1-4), 263-279.

Westra, S., and S. A. Sisson (2011), Detection of non-stationarity in precipitation extremes using a max-stable process model, *Journal of Hydrology*, *406*(1-2), 119–128.

Westerberg, I., J.-L. Guerrero, J. Seibert, K.J. Beven, and S. Halldin (2011), Stage-discharge uncertainty derived with a non-stationary rating curve in the Choluteca River, Honduras. *Hydrological Processes*, *25*, 603-613, doi: 10.1002/hyp.7848.

Westra, S., L. V. Alexander, and F. W. Zwiers (2013), Global increasing trends in annual maximum daily precipitation, *Journal of Climate*, *26*, 3904–3918.

Wilby, R., B. Greenfield, and C. Glenny (1994), A coupled synoptichydrological model for climate change impact assessment, *J. Hydrol.*, 153, 265–290.

Xu, C.-y., L. Gong, T. Jiang, D. Chen, and V. P. Singh (2006), Analysis of spatial distribution and temporal trend of reference evapotranspiration and pan evaporation in Changjiang (Yangtze River) catchment, *Journal of Hydrology*, *327*(1-2), 81–93.

Yadav, M., T., Wagener and H. Gupta (2007), Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins, *Advances in Water Resources*, *30*, 1756–1774.

Yaeger, M., E. Coopersmith, S. Ye, L. Cheng, A. Viglione, and M. Sivapalan (2012), Exploring the physical controls of regional patterns of flow duration curves Part 4: A synthesis of empirical analysis, process modeling and catchment classification, *Hydrol. Earth Syst. Sci.*, *16*, 44834498, doi:10.5194/hess-16-4483- 2012.

Yapo, P.O., H.V. Gupta, and S. Sorooshian (1996), Calibration of conceptual rainfall-runoff models: Sensitivity to calibration data, *Journal of Hydrology*, *181*, 23–48.

Yapo, P.O., H.V. Gupta, and S. Sorooshian (1998), Multi-objective global optimization for hydrologic models, *Journal of Hydrology*, *204*, 83–97.

Ye, S., M. Yaeger, E. Coopersmith, L. Cheng, and M. Sivapalan (2012), Exploring the physical controls of regional patterns of flow duration curves Part 2: Role of seasonality, the regime curve, and associated process controls, *Hydrol. Earth Syst. Sci.*, *16*, 4447–4465, doi:10.5194/hess-16-4447-2012.

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 Resources Research, 44*, W09417, doi:10.1029/2007WR006716.

Yokoo, Y. and M. Sivapalan (2011), Towards reconstruction of the flow duration curve: development of a conceptual framework with a physical basis, *Hydrol. Earth Syst. Sci.*, *15*, 28052819, doi:10.5194/hess-15-2805-2011.

Young, P.C. (2002), Advances in real-time flood-forecasting forecasting, *Philosophical Transactions Royal Society, Physical and Engineering Sciences*, *360*(9), 1433–1450.

Young, P. C. (2013), Hypothetico-inductive data-based mechanistic modeling of hydrological systems, *Water Resources Research*, *49*, doi:10.1002/wrcr.20068.

Yu, P.S., and T.C. Yang (1996), Synthetic regional flow duration curve for southern Taiwan, *Hydrolog. Process.*, *10*, 373391.

Yu, P.S. and T.C. Yang (2000), Using synthetic flow duration curves for rainfall-runoff model calibration at ungauged sites, *Hydrol. Process.*, *14*, 117–133.

Yu, P.S., T.C. Yang, and Y.C. Wang (2002), Uncertainty analysis of regional flow duration curves, *J. Water Resour. Pl. Manage.*, *128*(6), 424–430.

Yue, S., P. Pilon, G. Cavadias (2002), Power of the Mann-Kendall and Spearman's rho tests for detecting monotonic trends in hydrological series, *Journal of Hydrology*, *264*(14), 262–263.

Zhang, X., K. D. Harvey, W. D. Hogg, T. R. Yuzyk (2001), Trends in Canadian streamflow, *Water Resources Research 37*(4), 987-998.

Zhang, Y. K., and K. E. Schilling (2006), Increasing streamflow and baseflow in Mississippi River since the 1940s: Effect of land use change, *Journal of Hydrology*, *324*, 412-422.

Zhang, L., N. Potter, K. Hickel, Y. Zhang, and Q. Shao (2008), Water balance modeling over variable time scales based on the Budyko framework Model development and testing, *J. Hydrol.*, *360*, 117131, doi:10.1016/j.jhydrol.2008.07.021.

Zhang, Q., C.-y. Xu, S. Becker, Z. X. Zhang, Y. D. Chen, and M. Coulibaly (2009), Trends and abrupt changes of precipitation maxima in the Pearl River basin, China, *Atmospheric Science Letters*, *10*, 132–144. doi: 10.1002/asl.221

Zhao, F., Z. Xu and L. Zhang (2012), Changes in streamflow regime following vegetation changes from paired catchments, *Hydrol. Process.*, *26*, 1561–1573, doi:10.1002/hyp.8266.