Ancient numerical daemons of conceptual hydrological modeling: 2. Impact of time stepping schemes on model analysis and prediction

Dmitri Kavetski[1] and Martyn P. Clark[2]

Received 12 November 2009; revised 19 March 2010; accepted 21 April 2010; published 8 October 2010.

[1] Despite the widespread use of conceptual hydrological models in environmental research and operations, they remain frequently implemented using numerically unreliable methods. This paper considers the impact of the time stepping scheme on model analysis (sensitivity analysis, parameter optimization, and Markov chain Monte Carlo-based uncertainty estimation) and prediction. It builds on the companion paper (Clark and Kavetski, 2010), which focused on numerical accuracy, fidelity, and computational efficiency. Empirical and theoretical analysis of eight distinct time stepping schemes for six different hydrological models in 13 diverse basins demonstrates several critical conclusions. (1) Unreliable time stepping schemes, in particular, fixed-step explicit methods, suffer from troublesome numerical artifacts that severely deform the objective function of the model. These deformations are not rare isolated instances but can arise in any model structure, in any catchment, and under common hydroclimatic conditions. (2) Sensitivity analysis can be severely contaminated by numerical errors, often to the extent that it becomes dominated by the sensitivity of truncation errors rather than the model equations. (3) Robust time stepping schemes generally produce "better behaved" objective functions, free of spurious local optima, and with sufficient numerical continuity to permit parameter optimization using efficient quasi Newton methods. When implemented within a multistart framework, modern Newton-type optimizers are robust even when started far from the optima and provide valuable diagnostic insights not directly available from evolutionary global optimizers. (4) Unreliable time stepping schemes lead to inconsistent and biased inferences of the model parameters and internal states. (5) Even when interactions between hydrological parameters and numerical errors provide "the right result for the wrong reason" and the calibrated model performance appears adequate, unreliable time stepping schemes make the model unnecessarily fragile in predictive mode, undermining validation assessments and operational use. Erroneous or misleading conclusions of model analysis and prediction arising from numerical artifacts in hydrological models are intolerable, especially given that robust numerics are accepted as mainstream in other areas of science and engineering. We hope that the vivid empirical findings will encourage the conceptual hydrological community to close its Pandora's box of numerical problems, paving the way for more meaningful model application and interpretation.

Citation: 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 Resour. Res.*, *46*, W10511, doi:10.1029/2009WR008896.

1. Introduction

[2] Hydrological models are routinely used in both scientific and operational contexts. For example, in operational contexts they are used in flood forecasting, water resource assessments, and other environmental management, whereas in scientific studies they can help understand hydrological processes. Of particular significance in both contexts are computationally fast conceptual hydrological models, which may capture key catchment dynamics given only limited environmental data. The parameters of hydrological models are determined using a combination of prior knowledge (from previous studies and/or similar catchments) and calibration to observed forcing-response data (typically, rainfall-runoff). A priori parameter estimation is challenging for several related reasons, including (1) soil properties and vegetation characteristics have tremendous spatial variability, both within and between basins [*Miller and White*, 1999]; and (2) it is extremely difficult to relate conceptual model parameters to the available spatial information on soils and vegetation [*Koren et al.*, 2003; *Duan et al.*, 2006]. The parameter values and distributions are therefore usually inferred using inverse methods (i.e., calibration), which seek to identify parameter

[1]Environmental Engineering, University of Newcastle, Callaghan, New South Wales, Australia.
[2]National Center for Atmospheric Research, Boulder, Colorado, USA.

0043-1397/10/2009WR008896

set(s) that provide the best "fit" to the observed data, quantified using an objective function.

[3] Reliable and efficient calibration of conceptual hydrological models has been a major research and practical challenge over the last three decades [e.g., *Beven and Binley*, 1992; *Kavetski et al.*, 2003a; *Vrugt et al.*, 2008; *Renard et al.*, 2010, and many others]. Apart from the difficulty in selecting an appropriate objective function, many difficulties also arise due to its geometrical complexity. As demonstrated by *Duan et al.* [1992], macroscale local optima, in collusion with macro and micro discontinuities, undermine traditional locally convergent gradient-based Newton-type optimization. These problems have influenced several major directions of hydrological calibration research over the last decades, including a major focus on algorithms to handle multimodality and nonsmoothness in objective functions [e.g., *Duan et al.*, 1992; *Thyer et al.*, 1999; *Tolson and Shoemaker*, 2007] and encouraging calibration paradigms less reliant on optimization and well-behaved objective functions [e.g., *Beven and Binley*, 1992].

[4] Despite recent findings that, in many cases, these troublesome difficulties are avoidable numerical artifacts [*Kavetski et al.*, 2003b, 2006a], the numerical implementation has remained a neglected weakness of hydrological models. In particular, the companion paper [*Clark and Kavetski*, 2010] surveyed the spectrum of conceptual and physically motivated hydrological models, both lumped, semidistributed and distributed, and found a prevalence of fixed-step explicit time stepping approximations. This is despite error control being widely recognized as essential for reliable numerical computing [e.g., *Press et al.*, 1992; *Shampine and Reichelt*, 1997] and despite growing indications in the hydrological literature that poor model numerics have a detrimental effect on model performance and application [e.g., *Kavetski et al.*, 2003b; *Kavetski and Kuczera*, 2007].

[5] This 2-part study provides a broad assessment of the impact of numerical approximation on key aspects of model development and application. The first paper [*Clark and Kavetski*, 2010] evaluated several classes of time stepping schemes for conceptual hydrological models, with a focus on numerical accuracy, fidelity in the context of fitting observed data, and computational efficiency. It vividly showed that, in many cases, the numerical errors of uncontrolled time stepping schemes clearly exceed likely model errors arising from structural simplifications in the governing equations and data errors arising from observational uncertainty. Conversely, numerical error control (adaptive explicit approximations) or reliance on unconditional stability (fixed-step implicit solutions) efficiently addresses these concerns. The analysis was carried out for eight time stepping approximations of six distinct models of varying degree of complexity, applied to 13 basins with different hydroclimatic and physical conditions, and hence provides strong and broad empirical evidence supporting similar findings in previous work [e.g., *Kavetski et al.*, 2003b, 2006a].

[6] In this paper we consider the impact of the time stepping scheme on the results and conclusions of model application and predictions, including parameter sensitivity analysis and optimization, as well as model inference and validation. The significance of these steps is elaborated below.

[7] Sensitivity analysis yields key insights into model behavior and helps identify redundant or poorly behaved parameters and model components (e.g., see *Saltelli* [2002] for theory and *Wagener et al.* [2009] for hydrological applications). Given the troublesome macroscale numerical artifacts in model objective functions reported by *Kavetski et al.* [2003b], we investigate the contamination of parameter sensitivity estimates by numerical errors of the time stepping schemes. Despite a widespread use of sensitivity analysis in hydrology [e.g., *Pappenberger et al.*, 2008; *van Werkhoven et al.*, 2008; *Yatheendradas et al.*, 2008], we are not aware of any previous assessments of its robustness with respect to numerical approximation errors in the model equations.

[8] Parameter optimization is critical in model calibration and application. Numerous optimization methods have been used in conceptual hydrological modeling, with evolutionary-type methods increasingly dominating research and practice [e.g., *Duan et al.*, 1992; *Thyer et al.*, 1999; *Vrugt and Robinson*, 2007]. In addition to operational use, calibrated parameters are frequently used to understand model behavior and interpret catchment dynamics [*Ivanov et al.*, 2004] and to estimate regional parameter values, e.g., for prediction in ungauged basins [e.g., *Merz and Blöschl*, 2004]. Given indications that model distortions arising from time stepping errors affect both the optimal parameters themselves and the ability to estimate these parameters using optimization algorithms [*Kavetski et al.*, 2003b, 2006b], this paper evaluates and discuses the performance of modern gradient-based and evolutionary optimization methods with respect to time stepping approximations in different model structures applied over multiple catchments.

[9] Finally, given significant uncertainties in the hydrological data and the structural errors in the current generation of hydrological models, rigorous uncertainty assessment is increasingly viewed as essential in hydrological calibration and prediction (e.g., see *Beven* [2008] for a thorough discussion). This paper explores the effects of numerical time stepping errors on the conclusions reached from uncertainty analysis, including the inferred distributions of model parameters and internal states. The robustness of model predictions, especially in validation mode, is critical for meaningful model application and is also examined in this study. Since current practice is progressively dominated by Markov chain Monte Carlo (MCMC) methods [e.g., *Kuczera and Parent*, 1998; *Vrugt et al.*, 2009], we also provide some comments on the effect of time stepping artifacts on the convergence of MCMC algorithms.

[10] The broader evaluation of the impact of numerical model errors on sensitivity analysis, parameter optimization, uncertainty assessment, and model interpretation significantly advances previous work in this direction [*Kavetski et al.*, 2003b, 2006a, 2006b]. Importantly, the inclusion of several distinct classes of time stepping schemes, a range of conceptual hydrological models of varying complexity and multiple basins with diverse physical and hydroclimatic characteristics provides much stronger and broader evidence of the generality and pertinence of our conclusions.

[11] The paper is organized as follows. Sections 2 and 3 describe, respectively, the models and numerical time stepping schemes used in the empirical evaluation, while section 4 describes the basins and hydrological data. Section 5 provides details of the methods used for sensitivity analysis, optimization, and parameter inference and prediction. Section 6 provides an analysis of objective function surfaces, while sections 7, 8 and 9 detail, respectively, impacts of

numerical approximation errors on model sensitivity, optimization, and prediction. Drawing on these results and those of the companion paper, section 10 provides practical guidance on the selection of time stepping methods for conceptual hydrological models. Finally, section 11 summarizes the major findings and discusses the broad implications of this study for the discipline of hydrology.

2. Hydrological Models

[12] The impact of numerical implementation on model analysis and interpretation is examined using six models from the Framework for Understanding Structural Errors (FUSE) hydrological tool kit [*Clark et al.*, 2008]. The selected models, FUSE-070, FUSE-060, FUSE-536, FUSE-550, FUSE-092, and FUSE-330, are broadly representative of the wide spectrum of conceptual hydrological models used in research and practice. They include configurations where the unsaturated zone has one or two state variables (e.g., tension and free storage), models with linear and nonlinear base flow and models where surface runoff is represented as a function of either upper or lower layer storages. See the companion paper [*Clark and Kavetski*, 2010] for further details.

[13] All FUSE models are formulated in state-space form as ordinary differential equations (ODEs),

$$\frac{d\mathbf{S}}{dt} = \mathbf{g}(\mathbf{S}), \tag{1}$$

where $\mathbf{S}$ represents storage in the various conceptual compartments of the model and $\mathbf{g}(\mathbf{S})$ is assembled using the hypothesized flux formulations and connectivities of the model stores.

[14] Most hydrological models can be cast as ODE systems such as equation (1). Since analytical solutions are unavailable except in highly simplistic cases, numerical approximations must be employed. The impact of these approximations on model analysis and application is the key focus of this paper.

3. Summary of Time Stepping Schemes

[15] This section outlines the time stepping schemes evaluated in this study. A detailed analysis of the numerical reliability and computational efficiency of these schemes can be found in the companion paper. The reader is also referred to classic texts on numerical ODE methods for further background material [e.g., *Lambert*, 1991; *Shampine*, 1994].

3.1. Explicit Methods

[16] The basic approach that remains prevalent in conceptual hydrological models is the explicit Euler scheme,

$$\mathbf{S}^{n+1}_{\mathrm{EE}(1)} = \mathbf{S}^n + \Delta t\, \mathbf{g}(\mathbf{S}^n). \tag{2}$$

The explicit Euler method is first order accurate: numerical errors in $\mathbf{S}^{n+1}_{\mathrm{EE}(1)}$ are $O(\Delta t)$, i.e., linearly proportional to the time step size Δt.

[17] The effects of increasing the order of accuracy are examined using the explicit Heun scheme,

$$\mathbf{S}^{n+1}_{\mathrm{EH}(2)} = \mathbf{S}^n + \frac{\Delta t}{2}\left[\mathbf{g}(\mathbf{S}^n) + \mathbf{g}\left(\mathbf{S}^{n+1}_{\mathrm{EE}(1)}\right)\right], \tag{3}$$

which uses the explicit Euler estimate (2) as an intermediate stage and yields $O(\Delta t^2)$ accuracy.

3.2. Implicit Methods

[18] The benefits of unconditional stability, as opposed to higher-order approximations or adaptive substepping, are explored using the first order implicit Euler scheme,

$$\mathbf{S}^{n+1}_{\mathrm{IE}(1)} = \mathbf{S}^n + \Delta t\, \mathbf{g}\left(\mathbf{S}^{n+1}_{\mathrm{IE}(1)}\right). \tag{4}$$

The implicit Heun (Crank-Nicholson) method provides $O(\Delta t^2)$ accuracy [e.g., *Wood*, 1990],

$$\mathbf{S}^{n+1}_{\mathrm{IH}(2)} = \mathbf{S}^n + \frac{\Delta t}{2}\left[\mathbf{g}(\mathbf{S}^n) + \mathbf{g}\left(\mathbf{S}^{n+1}_{\mathrm{IH}(2)}\right)\right]. \tag{5}$$

While the implicit algorithms (4) and (5) require iterative solution at every time step and therefore are markedly costlier per step than their explicit counterparts (2) and (3), in practice their unconditional stability provides much better reliability and efficiency than conditionally stable explicit schemes for large stepsizes and "stiff" equations (e.g., see *Shampine* [1994] for a thorough theoretical explanation, and the companion paper for an evaluation in hydrological contexts).

3.3. Semi-implicit Methods

[19] For many problems, such as weakly nonlinear ODEs, the semi-implicit Euler scheme retains most stability benefits of implicitness, while avoiding expensive multiple iterations at each time step. It is given by

$$\mathbf{S}^{n+1}_{\mathrm{SIE}(1)} = \mathbf{S}^n + \left[\mathbf{I} - \Delta t \frac{\partial \mathbf{g}(\mathbf{S}^n)}{\partial \mathbf{S}}\right]^{-1} \Delta t\, \mathbf{g}(\mathbf{S}^n), \tag{6}$$

where $\mathbf{I}$ is the identity matrix and $\partial\mathbf{g}/\partial\mathbf{S}$ is the ODE Jacobian (see *Clark and Kavetski* [2010] for further details).

3.4. Numerical Error Control and Adaptive Substepping

[20] Since the quality of a time stepping approximation depends strongly on the stepsize Δt, error control is universally recognized as essential in numerical integration [e.g., *Kahaner et al.*, 1989]. For example, *Shampine and Reichelt* [1997] do not even allow fixed-step integration in their Matlab tool kit, even for unconditionally stable algorithms (because stability merely avoids uncontrolled error growth and cannot guarantee actual numerical accuracy). Conversely, current hydrological practice remains largely dominated by fixed-step conditionally stable methods (see review by *Clark and Kavetski* [2010] for some exceptions).

[21] To provide a thorough and pertinent evaluation of time stepping schemes in conceptual hydrology, especially in the applied context of fitting observed data, this paper includes both conditionally stable explicit and unconditionally stable implicit schemes in its analysis, both in fixed-step and adaptive-substepping implementations, as detailed in the companion paper. Given the didactic objectives of this two-part paper, we use a basic implementation of the adaptive error-controlled methods used in more sophisticated numer-

ical ODE solvers [e.g., *Kahaner et al.*, 1989], in particular, estimating the truncation error by comparing two different numerical approximations, and subdividing "outer" data-resolution steps until each individual substep satisfies a mixed error test with absolute and relative truncation error tolerances τ_A and τ_R [see *Clark and Kavetski*, 2010, Appendix B]. In all cases, daily rainfall-runoff data is used; we refer to the companion paper for an investigation of the impact of the time scale of the numerical reliability and efficiency of the time stepping schemes. We also envisage applying more sophisticated variable-order solvers from canned numerical ODE packages in a subsequent paper.

4. Hydrological Data

4.1. Mahurangi (MARVEX) Data

[22] A detailed analysis is carried out using the hydroclimatic data from the Mahurangi River Variability Experiment (MARVEX) in Northland, New Zealand [see *Woods et al.*, 2001; *Ibbitt and Woods*, 2004]. In this study, we used daily basin average rainfall estimates obtained from 13 raingauges [see *Woods et al.*, 2001], potential evapotranspiration estimated from temperature, humidity, and solar radiation [*Tait and Woods*, 2007] and daily streamflow gauged at the Auckland Regional Council station at Mahurangi at College (drainage area is 46.65 km^2). The calibration length for this study was 1492 days, with the first 297 days used as a warmup.

4.2. MOPEX Data

[23] A broader but less detailed assessment is carried out using 12 catchments from the Model Parameter Estimation Experiment (MOPEX) [*Duan et al.*, 2006]. Here, the FUSE models are forced with combined rain and snowmelt estimates obtained from the SNOW-17 model used for the National Weather Service MOPEX simulations (i.e., using SNOW-17 as a pre-processor of the original MOPEX precipitation time series) [*Clark et al.*, 2008]. The MOPEX basins represent diverse hydroclimatic and land surface conditions, spanning dry to wet regimes, croplands to mixed forests, and a range of soil types (see Table 3 in the companion paper for further details). Eleven years of data (1980–1990) were used in calibration (using 1979 for model warmup), while 20 individual years in the 1960−1979 period were used in validation. Carrying out the calibration and validation studies over this diverse range of catchments and multiple time periods yields both broad and specific insights and asserts the generality of the conclusions.

5. Methodology

5.1. Examination of Objective Function Surfaces

[24] The objective function used in this work is the root-mean-squared error (RMSE) of streamflow predictions

$$\Phi_{\mathrm{RMSE}}[\boldsymbol{\theta}] = \sqrt{\frac{1}{N_t}\sum_{n=1}^{N_t}\left(q_{\mathrm{sim}}^{n}[\boldsymbol{\theta}] - q_{\mathrm{obs}}^{n}\right)^2}, \tag{7}$$

where q_{obs}^{n} is the observed streamflow at the nth time step, $q_{\mathrm{sim}}^{n}[\boldsymbol{\theta}]$ is the model streamflow produced using parameter set $\boldsymbol{\theta}$, and N_t is the total number of time steps.

[25] The Nash-Sutcliffe index (NS), commonly used in hydrology, is related to RMSE (7) via

$$\Phi_{\mathrm{NS}}[\boldsymbol{\theta}] = 1 - \left(\frac{\Phi_{\mathrm{RMSE}}[\boldsymbol{\theta}]}{\sigma_{\mathrm{obs}}}\right)^2, \tag{8}$$

where σ_{obs}^2 is the variance of the observations about their global mean.

5.2. Parameter Sensitivity Analysis

[26] Global parameter sensitivity is assessed using the *Saltelli* [2002] implementation of the Sobol' method [*Sobol'*, 1993]. The Sobol' method decomposes the total model sensitivity into contributions from individual parameters and their interactions (see Appendix A for details). While we restrict numerical experiments to the total global sensitivity of individual parameters, we also briefly comment on the implications for sensitivity analysis of their interactions.

[27] The total global sensitivity indices S_j^{TOT}, based on the RMSE model performance measure, were calculated using 10,000 parameter sets sampled using quasi-random Sobol' sequences [*Bratley and Fox*, 1988]. The same parameter sets were used to calculate S_j^{TOT} for all time stepping schemes, ensuring that any differences in S_j^{TOT} are due solely to differences in the model implementation.

5.3. Parameter Optimization

[28] This paper considers the impact of the time stepping scheme on two distinct strategies for parameter optimization: (1) a multistart quasi-Newton optimizer with finite difference derivatives [*Dennis and Schnabel*, 1996; *Nocedal and Wright*, 1999; *Kavetski et al.*, 2007] and (2) the Shuffled Complex Evolution (SCE) global optimizer [*Duan et al.*, 1992; *Thyer et al.*, 1999].

5.3.1. Quasi-Newton Optimization

[29] Quasi-Newton optimization constructs and updates a quadratic local approximation to the objective function to move toward the nearest local optimum [*Nocedal and Wright*, 1999]. We used a modern quasi-Newton code with (1) a trust-region method to stabilize convergence to the nearest local optimum [*Dennis and Schnabel*, 1996], (2) an active set method to efficiently handle parameter bounds by sliding along boundary subspaces [*Nocedal and Wright*, 1999], (3) adaptive finite difference gradient approximation, and (4) multiple independent starts randomly seeded through the parameter space to improve the probability of locating the global optimum and gain insights into the large-scale multimodality of the objective function [*Kavetski et al.*, 2007].

[30] The Newton optimizer used 100 random starting seeds, with scaled convergence tolerances of 10^{-10} for the parameters and the objective function and a maximum objective function evaluation count of 5000 for each local optimization seed. The same seeds were used for all time stepping schemes. The convergence tolerance was tight to avoid premature termination in near-flat regions of the objective function. We also note that the empirical analysis in this paper focuses primarily on the general qualitative performance of the optimizer rather than on a detailed quantitative evaluation of the influence of tolerances, algorithmic settings, etc.

5.3.2. Shuffled Complex Evolution (SCE) Search

[31] The SCE optimizer was developed to handle the discontinuous multimodal objective functions identified in hydrological calibration nearly 20 years ago [*Duan et al.*, 1992]. It evolves a population of function samples in the feasible parameter space, treating each sample as a vertex of a simplex [*Nelder and Mead*, 1965]. In addition to the elongations, reflections, and contractions used in standard simplex methods, the SCE algorithm periodically shuffles the vertices of different simplexes to exchange information. This prevents entrapment in local optima and enhances the global convergence of the algorithm. SCE is one of the most widely used optimization methods for conceptual models, attesting to the recognition of its robustness by the hydrological community.

[32] In line with *Duan et al.* [1994], the SCE optimizer used 10 complexes, with $2N_d + 1$ points in each complex, $N_d + 1$ points in each subcomplex and $2N_d + 1$ evolution steps before shuffling (where N_d is the dimensionality of the objective function, here equal to the number of model parameters). The SCE search was set to terminate if the objective function (here, the RMSE) did not decrease by more than 0.001 mm/d after nine shuffles, and the total number of objective function calls was limited to 100,000 (never reached). Since the chief focus of this paper is to demonstrate the impact of time stepping errors on model analysis and prediction, detailed effects of convergence and algorithmic settings are beyond its scope (we refer the reader to *Madsen et al.* [2002], *Tolson and Shoemaker* [2007], and *Behrangi et al.* [2008]).

5.4. Inference and Prediction

5.4.1. Background Theory

[33] Optimization of the RMSE criterion (7) is closely related to least squares regression commonly used in statistical estimation [*Box and Tiao*, 1992]. The latter corresponds to the Bayesian posterior distribution $p(\boldsymbol{\theta}|\mathbf{D})$ of parameters $\boldsymbol{\theta}$ given observed data $\mathbf{D}$ and prior information $p(\boldsymbol{\theta})$,

$$p\left(\boldsymbol{\theta}, \sigma_y^2 | \mathbf{D}\right) = L\left(\mathbf{D} \left| \boldsymbol{\theta}, \sigma_y^2 \right.\right) p\left(\boldsymbol{\theta}, \sigma_y^2\right), \tag{9}$$

where $\mathbf{D} = \{\tilde{\mathbf{x}}, \tilde{\mathbf{y}}\}$ comprises the observed forcing $\tilde{\mathbf{x}}$ and responses $\tilde{\mathbf{y}}$ of length N_t, and $L(\mathbf{D}|\boldsymbol{\theta}, \sigma_y^2)$ is the likelihood function. For ordinary least squares regression,

$$L\left(\mathbf{D} \left| \boldsymbol{\theta}, \sigma_y^2 \right.\right) = \prod_{n=1}^{N_t} N\left(\tilde{y}^n - h^n(\boldsymbol{\theta} \,|\tilde{\mathbf{x}}) \left| 0, \sigma_y^2 \right.\right), \tag{10}$$

i.e., the independent Gaussian probability density function $N(\mathrm{e}[\boldsymbol{\theta}]|0, \sigma_y^2)$ of model residuals $\mathbf{e} = \tilde{\mathbf{y}} - \mathbf{h}(\boldsymbol{\theta}|\tilde{\mathbf{x}})$ computed from the observed and modeled responses, $\tilde{\mathbf{y}}$ and $\mathbf{h}$, respectively, assuming zero mean and an unknown (and hence inferred) residual variance σ_y^2 [*Box and Tiao*, 1992].

[34] When noninformative priors $p(\boldsymbol{\theta})$ are used for all quantities of interest, maximization of the RMSE criterion is equivalent to determining the most likely Bayesian posterior estimates of $\boldsymbol{\theta}$ (not to be confused with the *expected* posterior estimates!). This also corresponds to the classical maximum-likelihood approach under the assumption of independent Gaussian residuals, which yields the likelihood function (10) [*Box and Tiao*, 1992].

[35] Generally speaking, the inference of parameter distributions is much more informative than mere optimization: in addition to estimating the parameter values, it provides information about their uncertainty, potential multimodality, etc. This can guide model improvement and data collection investments to reduce these uncertainties.

[36] While the assumptions underlying least squares regression are restrictive [e.g., *Beven and Binley*, 1992; *Kavetski et al.*, 2003a], this paper focuses on the impact of numerical time stepping errors on model performance, which is generally gauged using RMSE-type measures, and therefore, more complicated statistical inferences lie outside its scope. However, the insights from this work are already guiding the selection of numerical methods for hydrological models analyzed using more advanced, but also computationally costlier, inference approaches such as Bayesian Total Error Analysis (BATEA) [*Kavetski et al.*, 2006c]. We will report on the BATEA calibration of FUSE models to the Mahurangi catchment in a separate paper that exploits the insights gained in this numerical evaluation to avoid obscuring its results and conclusions by numerical artifacts.

5.4.2. MCMC Methods

[37] As the nonlinearity of the model $\mathbf{h}(\boldsymbol{\theta})$ with respect to its parameters increases, its posterior distribution becomes progressively non-Gaussian [*Bates and Watts*, 1988]. Maximizing such distributions and, more generally, determining and reporting their shape and characteristics, even as basic as mean and variance, is usually impossible using analytical techniques. Instead, they must be explored numerically, e.g., using Markov chain Monte Carlo (MCMC) methods that are increasingly popular in hydrological practice. MCMC methods offer considerable flexibility in adapting to the shape of the distribution they are applied to and, provided care is taken to monitor their convergence, can approximate probability distributions that are intractable using alternative approaches (e.g., see *Gelman et al.* [2003] for further theory and *Kuczera and Parent* [1998], *Bates and Campbell* [2001], and *Vrugt et al.* [2003] for hydrologic applications).

[38] The adaptive MCMC strategy used in this study is described by *Thyer et al.* [2009]. Initially, the jump distribution is tuned one-parameter-at-a-time to achieve adequate jump ratios (~25%) over the first 2000 samples. The jump covariance matrix is then computed from the samples, and during the next 2000 samples, further tuned using a multiplicative factor. Following this, the jump covariance is fixed (i.e., the jump distribution is no longer adapted) and 35,000 samples are collected. The first 25,000 samples are treated as a warmup period, and only the final 10,000 "production" samples are used to construct and examine the parameter distributions. In the majority of cases, the Gelman-Rubin convergence test [*Gelman et al.*, 2003] was close to unity, suggesting, though not proving, the convergence of the MCMC chains to the target distribution.

6. Analysis of Objective Function Surfaces

6.1. Two-Dimensional Contour Plots

[39] Figure 1 shows the Nash-Sutcliffe contours for a representative cut through the (b, $S_{1,\max}$) parameter subspace of FUSE-536 in the Mahurangi basin. Figure 1 illustrates three related problems. First, the fixed-step explicit Euler and explicit Heun schemes dramatically degrade the model performance over parameter regions that produce good simula-

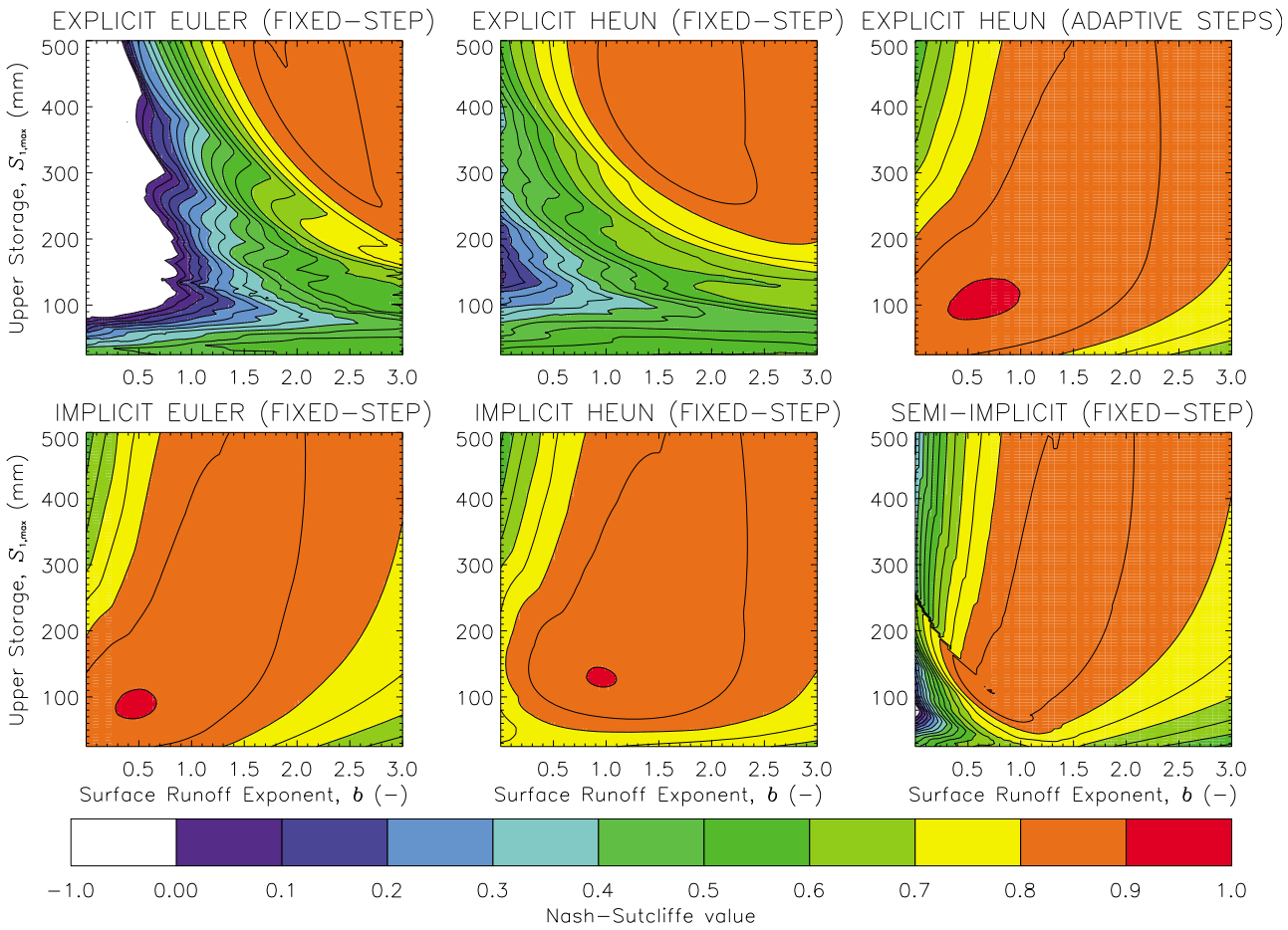

Figure 1. Impact of time stepping schemes on the objective function surface of conceptual hydrological models. Here, representative 2-D slices through the (b, $S_{1,\max}$) subspace of the Nash-Sutcliffe objective function of FUSE-536 applied to the Mahurangi basin are shown, with all other model parameters held constant at the same values. The numerical artifacts of uncontrolled time stepping are evident: The objective function of the underlying governing equations is remarkably well behaved, yet the objective functions of the same equations solved using fixed-step explicit methods are afflicted by massive macroscale deformations and extensive microscale noise. The semi-implicit approximation is also visibly deficient.

tions when the governing model equations are solved accurately (in particular, using adaptive error control). Second, the objective function surface has substantial macroscale roughness in the fixed-step explicit Euler and explicit Heun schemes, and, to a lesser extent, in the fixed-step semi-implicit scheme. Third, the near-optimal parameter values of the hydrological models solved using the fixed-step explicit Euler and explicit Heun schemes are different than those corresponding to the other time stepping schemes.

[40] The problems with macroscale distortions and macroscale roughness in fixed-step explicit schemes considerably complicate model calibration. Indeed, precisely as a result of such problems, the hydrological community has devoted considerable effort to developing calibration algorithms and paradigms that do not rely on well-behaved objective functions [*Beven and Binley*, 1992; *Duan et al.*, 1992; *Tolson and Shoemaker*, 2007]. We will return to these issues in section 8.

[41] The potential to confound parameter inference is especially worrying. For example, the fixed-step explicit Euler scheme performs poorly for low values of the surface runoff exponent b because this causes significant additional infiltration and hence rapid filling/drainage of the upper soil store that cannot be reliably handled using the fixed-step explicit scheme. Indeed, the "distortion ridge" along $S_{1,\max}$ = 100 mm in the fixed-step explicit Euler scheme (Figure 1) arises due to large fractional storage within storm events, which generates huge interflow and drainage fluxes when uncontrolled explicit schemes are used (see companion paper for further discussion). Note that accurate solutions of the governing model equations (adaptive explicit Heun scheme) or even just fixed-step unconditionally stable approximations (fixed-step implicit Euler scheme) have well-behaved Nash-Sutcliffe profiles. This implies that the striking differences between Figures 1a−1f are due entirely to numerical artifacts of uncontrolled explicit time stepping.

[42] Solely gridding the objective function cannot (by itself) indicate whether the "best" parameters are consistent with the process conceptualization. Nevertheless, numerically induced degradation in model performance over vast regions of the parameter space, e.g., for low values of exponent b for models implemented using the fixed-step explicit Euler scheme, is patently undesirable and can easily thwart efforts in process-based parameter analysis [*Gupta et al.*,

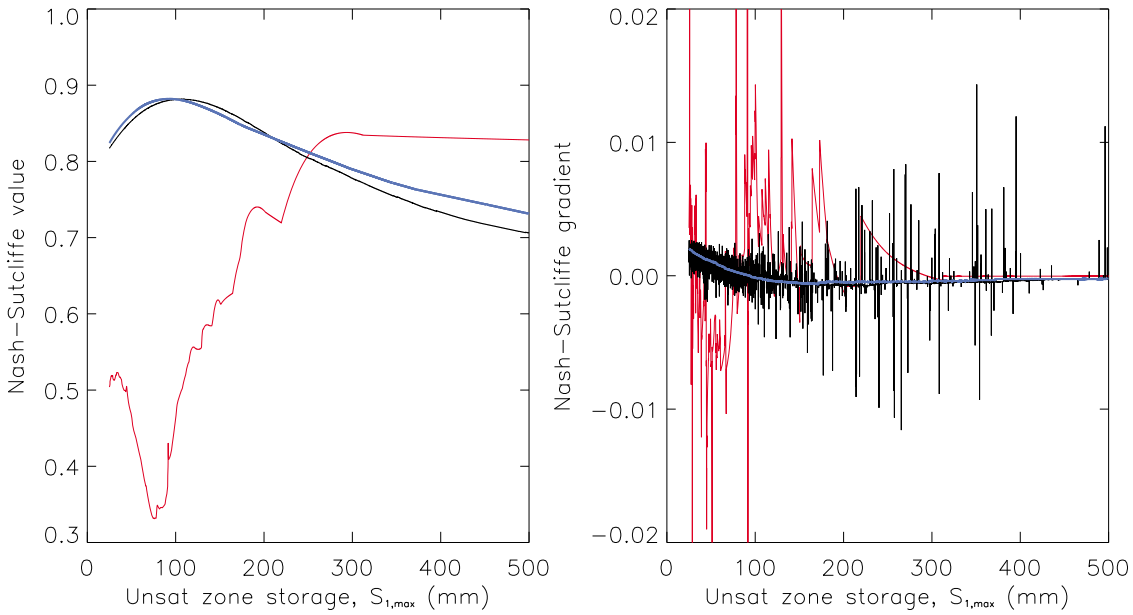

Figure 2. Representative slices through the Nash-Sutcliffe objective function and its gradient with respect to parameter $S_{1,\max}$ in FUSE-536 applied to the Mahurangi basin, for different time stepping methods (red = fixed-step explicit Euler scheme, blue = fixed-step implicit Euler scheme, and black = adaptive explicit Heun scheme with τ_R = 1% and τ_A = 0.01 mm). The surface runoff exponent b is set to 2.5 for the fixed-step explicit Euler and to 0.5 for the fixed-step implicit Euler and the adaptive explicit Heun; all other model parameters are held constant at the same representative values as used in Figure 1. The gradient $\partial\Phi_{NS}/\partial S_{1,\max}$ is estimated using one-sided finite differences ($\Delta S_{1,\max} = 0.475 \times 10^{-4}$ mm). Results show that while (left) both the fixed-step implicit and adaptive explicit solutions are free of macroscale deformations, (right) gradient analysis indicates that only the fixed-step implicit scheme is also free from microscale noise (see also Figure 3).

2008]. Section 9 further considers the impact of the time stepping scheme on both the parameter inference and actual model predictions, including internal states.

6.2. One-Dimensional Cross Sections

[43] To provide a more detailed inspection of the objective function, Figure 2 shows cross sections of the Nash-Sutcliffe index and its gradient with respect to parameter $S_{1,\max}$. As expected from Figure 1, widespread macroscale distortions and macroscale roughness occur when the model equations are solved using the fixed-step explicit Euler scheme. Moreover, analysis of the objective function gradient (Figure 2) also exposes microscale roughness in the adaptive explicit Heun method that is difficult to detect from an inspection of the Nash-Sutcliffe profile itself.

[44] Zooming in on the Nash-Sutcliffe profile, as illustrated in Figure 3, reveals two distinct patterns of microscale behavior of the adaptive explicit Heun solution: (1) seemingly random high-frequency microscale noise for low values of $S_{1,\max}$ and (2) more rare episodic (though still "periodic") "slips" at higher $S_{1,\max}$ values. Both pathologies are caused by changes in the number of substeps needed to meet the (fixed) truncation error tolerance when adaptive error control is enforced (see the authoritative discussion by *Gill et al.* [1981] and an earlier illustration by *Kavetski et al.* [2006a] using a single-state Variable Infiltration Capacity (VIC)-type model).

[45] As seen in Figure 3, the microscale noise produces small "pits" (local optima) in the objective function. Such pits can prevent a gradient-based optimization scheme from finding uphill directions and checking for convergence based on vanishing gradients [*Gill et al.*, 1981]. The more isolated episodic slips are less problematic, while they will certainly corrupt finite difference gradient estimates straddling such discontinuities, they do not appear to create spurious local optima, and, moreover, appear quite rare. In this work, they were observed in less than 100 of the 10,000 parameter sets forming the profile shown in Figure 2, i.e., in less than 1% of the parameter space. However, in general, their occurrence and frequency is probably case-specific and depends on the mathematical structure of the governing ODEs representing the model conceptualization, as well as on the forcing data, model parameters, etc. The impact of macro and microscale characteristics of the time stepping scheme on the model optimization is taken up in section 8.

7. Impact on Sensitivity Analysis

[46] Figure 4 shows the total global parameter sensitivity of all parameters of all FUSE models applied to the Mahurangi basin, estimated using the Sobol-Saltelli method (section 5.2). Several important insights can be drawn.

[47] 1. Models employing adaptive substepping have very similar sensitivities. This is unsurprising: The imposition of a 1% error tolerance makes numerical approximations effectively indistinguishable from the exact solution, especially in the context of parameter sensitivity analysis.

[48] 2. The fixed-step explicit Euler and explicit Heun schemes (and, to a lesser extent, the fixed-step implicit Heun

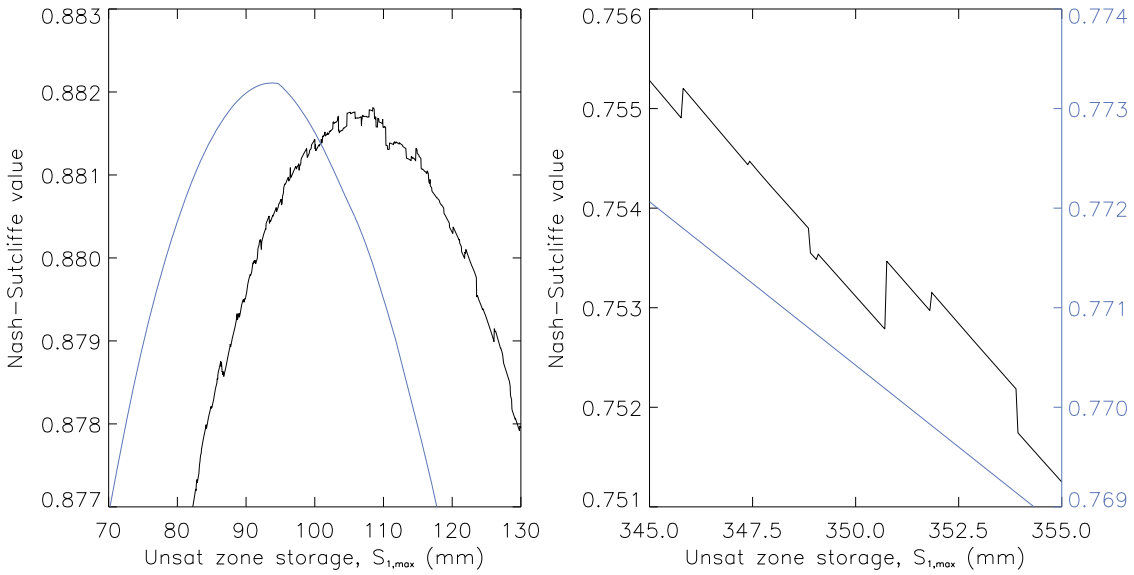

Figure 3. Representative 1-D slices through the Nash-Sutcliffe objective function with respect to parameter $S_{1,\max}$ in FUSE-536 applied to the Mahurangi basin, for the fixed-step implicit Euler (blue) and the adaptive explicit Heun (black), using the same parameters as in Figure 2. To contrast the microscale properties of both profiles in the zoomed right frame, a secondary *y* axis is used for the implicit Euler results. The microscale noise arising from adaptive substepping is evident, approximately of the order of the temporal truncation error tolerance τ_R = 1%. Fixed-step implicit time stepping provides much smoother model predictions, with microscale noise of the order of the Newton-Raphson iteration tolerance $\tau_N = 10^{-9}$ mm. Note that enforcing tight iteration tolerances τ_N generally does not require as much additional computational effort as tight temporal truncation error tolerances τ_A and τ_R [e.g., *Kavetski et al.*, 2006a].

and semi-implicit Euler schemes) have very different sensitivities, both from each other and from the adaptive methods. For example, the interflow rate k_i in FUSE-536 and FUSE-550 has a much higher sensitivity in the fixed-step explicit Euler and explicit Heun approximations than in the adaptive solutions. This finding is consistent with the RMSE analysis in Figures 3 and 4 of the companion paper, where fixed-step explicit methods were shown to be infidelious and unreliable. An erratic solution is likely to have a very different (indeed, erratic!) sensitivity than the exact solution, and this is precisely what is seen in Figure 4.

[49] 3. In almost all cases, the implicit Euler approximation has a similar sensitivity as the adaptive solutions, the main exception being parameter ϕ_{tens} in FUSE-330. This is a consequence of the high fidelity of the implicit Euler scheme and suggests that its accuracy is adequate for the purposes of sensitivity analysis (at least for the 6 models and 13 catchments considered in this study).

[50] To gauge the generality of the results in Figure 4, Figure 5 compares the parameter sensitivity of models implemented using the fixed-step explicit and implicit Euler approximations to the parameter sensitivity of the adaptive explicit Heun scheme (which is indistinguishable from the sensitivity of the underlying governing equations), for all 12 MOPEX basins.

[51] The fixed-step implicit Euler approximation, while inexact in the strict numerical accuracy sense, faithfully approximates the parameter sensitivities of the governing equations. In contrast, the sensitivity estimates for the fixed-step explicit Euler scheme are markedly corrupt. Indeed, Figures 4 and 5 indicate that a large fraction of parameter sensitivity in the explicit Euler scheme can be attributed to erratic behavior of this time stepping scheme, to the extent that, when viewed in conjunction with Figure 1 of the companion paper, the sensitivity analysis is effectively measuring the sensitivity of truncation errors rather than the sensitivity of the model itself!

[52] Finally, while this paper focused on the sensitivity of individual model parameters, similar findings are expect to apply to joint sensitivity analysis. Indeed, if the sensitivity of individual parameters is severely corrupted, it would be rather optimistic to expect their joint sensitivity to be preserved.

8. Impact on Model Optimization

8.1. Multi-start Quasi-Newton Method

[53] Macroscale distortions also affect model optimization. Figure 6 shows the termination points of multiple quasi-Newton optimization sequences applied to the RMSE objective function. It shows that, when the model is implemented using fixed-step explicit schemes, the quasi-Newton method frequently terminates at objective function values that are much lower than in the implicit Euler and adaptive schemes. This is especially true in FUSE-060, FUSE-536, and FUSE-550. A more detailed analysis indicated that most failures to reach the global optima were caused by lack of progress (inability to improve the objective function) rather than convergence to a "genuine" local optimum (where the gradient vanishes). These results are unsurprising given the considerable macroscale and microscale noise in the objective function that can be introduced by fixed-step explicit methods (Figure 1). It is precisely these features that motivated the abandonment of gradient-based methods in favor of

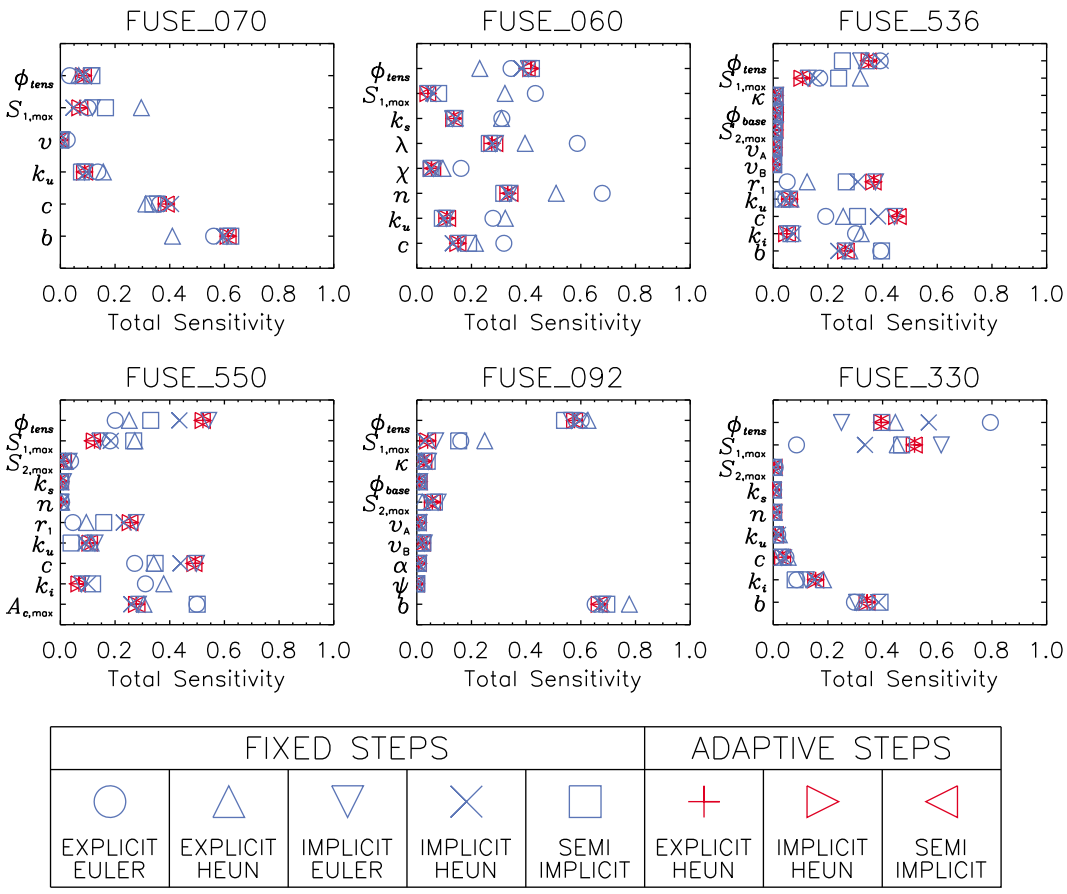

Figure 4. Impact of the time stepping scheme on the estimates of total global parameter sensitivity, obtained from 10,000 quasi-random Sobol samples, for all FUSE models in the Mahurangi basin. See *Clark and Kavetski* [2010] for parameter definitions. The parameter sensitivity analysis of models implemented using fixed-step explicit schemes are wildly different from those obtained with adaptive time stepping, indicating that the sensitivity analysis is dominated by truncation errors of the unreliable time stepping schemes.

global evolutionary methods [*Hendrickson et al.*, 1988; *Duan et al.*, 1992].

[54] Figure 6 also suggests that removal of the spurious distortions and increased smoothness of the objective functions dramatically improves the global convergence of the multi-start quasi-Newton method. The difference in parameter optimization of models with respect to the time stepping scheme was particularly pronounced for FUSE-060, where 70% of runs converged to near-global optimum values when the model was implemented using the fixed-step implicit Euler scheme, versus <5% when the same model was solved using the fixed-step explicit Euler or Heun schemes.

[55] However, even an accurate solution of the model equations does not guarantee a unimodal parameter distribution (see section 5.3.1). For example, in FUSE-060 and FUSE-536 solved using the implicit Euler scheme, up to ~35% (FUSE-060) and ~10% (FUSE-536) of the quasi-Newton sequences terminated in local optima. First, this confirms that a robust numerical implementation eliminates spurious multimodality but cannot possibly eliminate genuine multimodality that can arise when the governing model equations are strongly nonlinear with respect to their parameters, especially if the calibration data are also highly inaccurate [e.g., *Demidenko*, 2000; *Kavetski et al.*, 2006a]. Second, while traditionally viewed as a disadvantage, local convergence of individual quasi-Newton sequences to their nearest optima can be exploited to diagnose genuine multimodality. These aspects are examined further in section 8.4.

8.2. Comparison With a Global Optimizer: SCE Search

[56] The removal of spurious multimodality and improved smoothness of objective functions invites revisiting the comparison of reliability, informativeness, and computational efficiency of gradient-based algorithms versus evolutionary searches.

[57] To this end, Figure 6 compares the local optima reached by the multi-start quasi-Newton method with the global optima identified with SCE search (see section 5.3.2 for methodological details). Several important observations can be made.

[58] 1. The SCE method converges to similar objective function values for all the time stepping implementations of the same model. Algorithmically, this confirms the robustness of the SCE search as a global optimizer. From the point of view of model analysis, this further confirms that while fixed-step explicit schemes introduce gross errors over considerable parameter regions, there remain parameter sets for

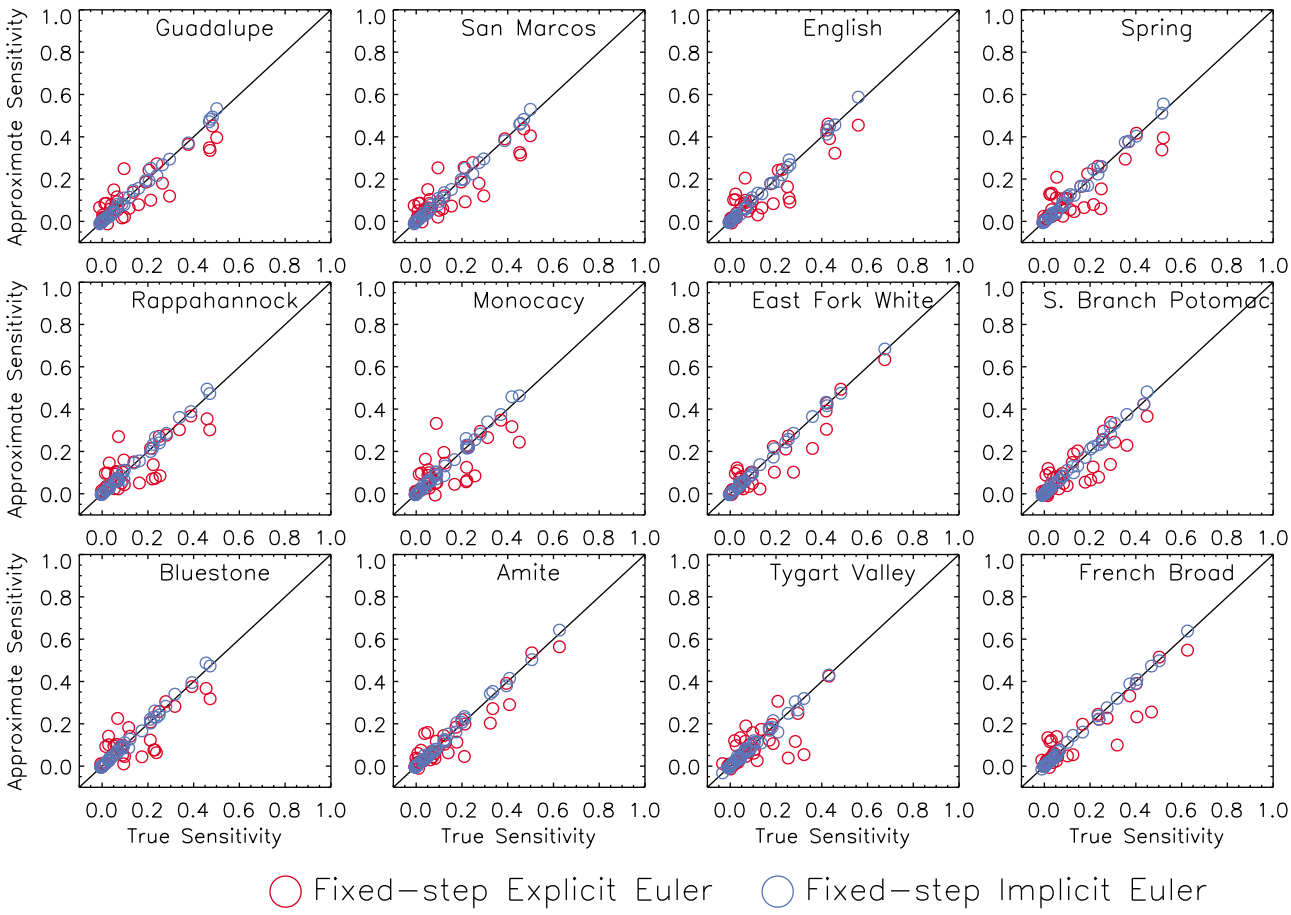

Figure 5. Broad comparison of the total global parameter sensitivity of models implemented using the fixed-step explicit Euler and fixed-step implicit Euler schemes versus the parameter sensitivity of the underlying model equations. The sensitivity of all parameters of all 6 FUSE models in all 12 MOPEX basins is shown. The analysis is based on 10,000 parameter sets sampled using the quasi-random Sobol' method.

which models implemented using fixed-step explicit schemes are competitive with the exact solution (e.g., see the RMSE scatterplots in the companion paper). This is examined in greater depth in section 9.3.

[59] 2. While the quasi-Newton method frequently terminates at lower objective function values when the hydrological models are implemented using the fixed-step explicit Euler and explicit Heun schemes, the best optima are nevertheless equivalent to those identified with SCE search. This shows that models implemented using robust numerical time stepping schemes, be it unconditionally stable fixed-step implicit approximations or adaptive (explicit or implicit) solutions, can be quite reliably calibrated using Newton-type optimizers. Provided the objective function is not exceedingly contaminated by spurious optima and discontinuities, Newton-type optimizers are competitive with SCE searches in reaching global optima.

[60] 3. In some rare cases (e.g., FUSE-092), the multi-start quasi-Newton method actually identifies parameter sets that have a slightly higher Nash-Sutcliffe index than that obtained by SCE. While not negating the robustness of the SCE search, this emphasizes the general competitiveness of Newton-type methods and is also demonstrative of the general inability of a single optimization algorithm to guarantee global solutions or provide the most efficient performance under all circumstances (e.g., see the "no free lunch" theorems of optimization [*Wolpert and Macready*, 1997]).

[61] The following sections elaborate on the computational efficiency and informativeness of multi-start quasi-Newton and SCE searches.

8.3. Speed of Progress Toward Solution

[62] Figure 7 compares the objective function values in the quasi-Newton and SCE methods for the first 1000 objective function evaluations (both the "best-so-far" and "current trial" values are shown). Since quasi-Newton optimization of inaccurate and/or exceedingly nonsmooth time stepping schemes is inadvisable (Figure 6), we show results for the fixed-step implicit Euler scheme.

[63] Figure 7 illustrates the advantages of exploiting the knowledge of uphill directions: The quasi-Newton sequences frequently converge to near-optimal objective function values within just 5–10 iterations, which, given the multiple objective function evaluations within each quasi-Newton iteration and the finite difference objective function gradient approximation, translates into ~200 model runs. The large "jumps" in the objective function values taken by the quasi-Newton method are noteworthy; they occur in near-quadratic

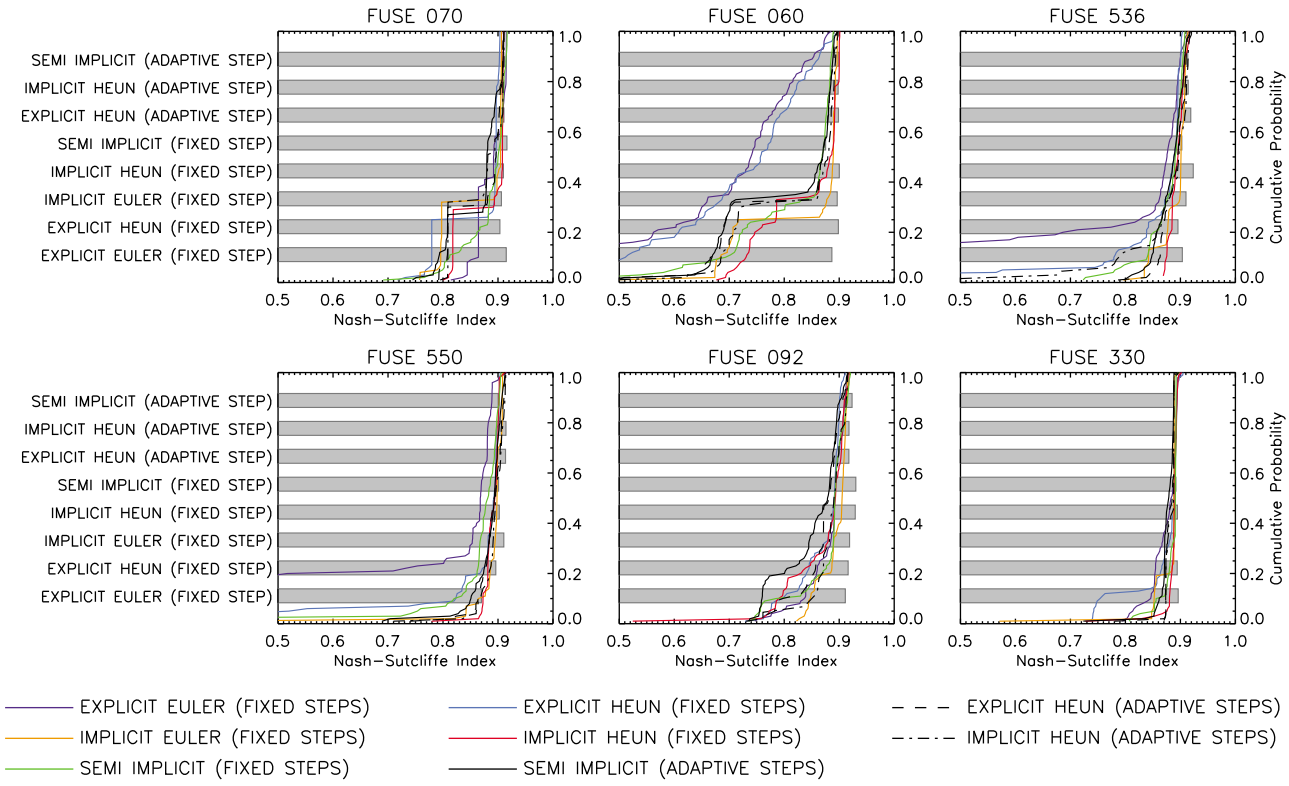

Figure 6. Parameter optimization of all FUSE models applied to the Mahurangi basin. The distributions of local optima estimated using the multi-start quasi-Newton method (colored lines) are compared with the single global optima estimated using the SCE search (gray bar). Results from 100 quasi-Newton sequences are shown. The cumulative distribution of optima identified by local optimization sequences indicates the fraction of the sequences converging to individual modes and is hence reflective of the probability mass associated with each distinct near-optimal parameter region. Pure global optimizers are less suited to this analysis.

regions of the parameter space, where the Newton equations can jump almost directly to the exact optimum.

[64] While the nonlinearity of the FUSE models makes their objective function nonquadratic (see analysis by *Kavetski et al.* [2006a]), the fast convergence even when starting far from the optimum solution suggests that a modern quasi-Newton optimizer, necessarily implemented within a multi-start framework, can robustly handle parameter calibration in realistic hydrological models. Indeed, it contradicts some literature [*Press et al.*, 1992] that suggests that the majority of the computational effort of Newton-type methods is spent getting close to the optimum, and once there, convergence is fast. For the problems in this study, we observed a different convergence behavior, with the quasi-Newton schemes quickly reaching near-optimal regions, followed by comparatively slow final convergence to the optimum.

[65] The robustness when far away from the optimum is due to the trust-region method incorporated into most modern Newton-type codes [*Conn et al.*, 2000], including the Matlab package [*Coleman et al.*, 2006]. The trust-region strategy limits the Newton correction to an adaptively updated region around the current estimate, where the current quadratic approximation of the objective function can be "trusted." In general, trust regions are more robust than linesearches and have strong convergence properties [*Conn et al.*, 2000].

[66] The "thrashing around" near the optimum appears to be caused by the microscale discontinuities in the objective function arising from adaptive substepping (see Figures 2 and 3). Consider that most optimization algorithms, perhaps excluding random search methods [e.g., *Tolson and Shoemaker*, 2007], estimate the optimal search direction from observed changes in objective function values. However, in near-optimal regions, where improvements in the objective function are necessarily small, discontinuities in the objective function necessarily interfere with the optimization algorithm: Changes in the objective values become contaminated, and eventually swamped, by numerical noise rather than genuine improvements. Quasi-Newton methods are particularly susceptible to this, since they compute the search direction from differences in function and gradient values over successive iterations [*Nocedal and Wright*, 1999]. As the optimum is approached, these differences become progressively dominated by numerical noise; indeed, differencing near-equal numerical values amplifies roundoff and other errors and is a well-known computational problem. Note that the simplex method underlying the SCE search is more robust but not immune in this respect: It relies on ranking the vertex function values to select the candidate for replacement. For any optimization algorithm, numerical noise fundamentally limits the precision to which the objective function can be optimized, and this should be reflected in the convergence tolerances (which in this study were set quite stringently for demonstration purposes).

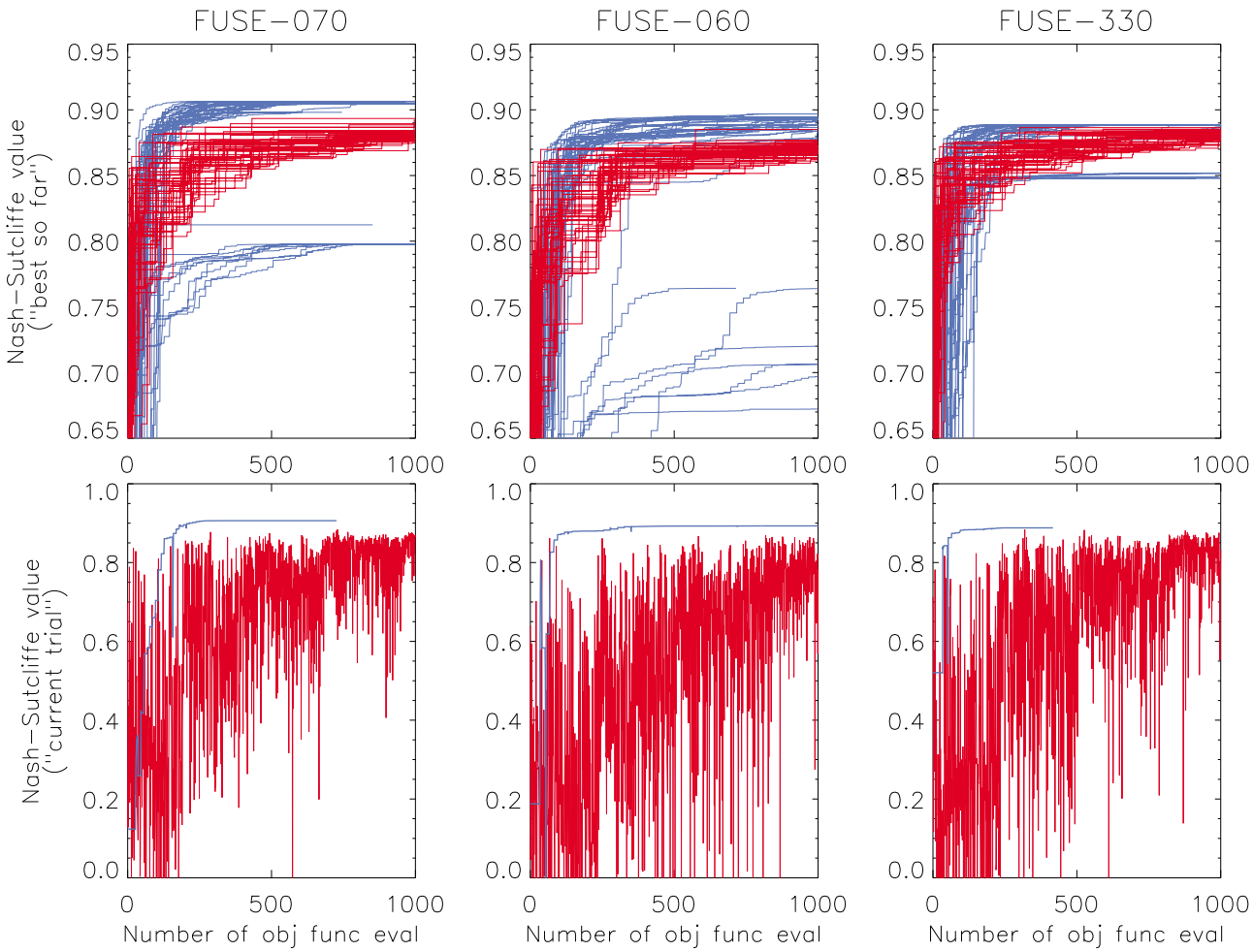

Figure 7. Representative comparison of the convergence of quasi-Newton (blue lines) and SCE optimizers (red lines), for several FUSE models implemented using the fixed-step implicit Euler approximation. (top) The "best-so-far" objective function values within 50 individual optimization sequences; (bottom) the trial values within a single typical optimization sequence. Each optimization sequence was initiated at a different point in the parameter space (obtained by quasi-random Sobol sampling). The first 1000 objective function evaluations within each optimization sequence are shown (note that several quasi-Newton sequences converge in fewer than 1000 function evaluations). The local convergence of quasi-Newton sequences is particularly evident for FUSE-060. Drops in function values during quasi-Newton optimization (bottom plot) represent function trials during trust region adaption and finite difference gradient approximation, and such drops are quite rare. Conversely, the SCE search continues to thoroughly explore clearly suboptimal regions of the search space, even when near-optimal regions are already found. This makes the SCE algorithm more robust but computationally slower. In addition, the use of derivative information by the quasi-Newton optimizer allows more rapid progress toward the nearest local optimum.

[67] Finally, Figure 7 also shows that the SCE algorithm, which relies on much less information to guide its progress toward the optimum and, moreover, repeatedly shuffles its complexes for a more thorough exploration of the parameter space, is slower than a Newton-type method when applied to smooth objective functions. The SCE search is still largely initializing and broadly exploring the parameter space in the first 200 function evaluations and does not begin to converge to near-optimal regions until after close to 1000 function calls. While altering the algorithm settings could improve the performance of any optimization method for a specific problem (e.g., see the exchange between *Behrangi et al.* [2008] and *Tolson and Shoemaker* [2008]), extensive theoretical analysis and empirical evidence [e.g., *Gill et al.*, 1981; *Nocedal and Wright*, 1999; *Conn et al.*, 2000] indicate that knowledge of the gradient and curvature of a smooth function permits an inherently faster progress toward the optimum than methods that do not exploit this information. Yet this only holds if the objective function is sufficiently smooth to be meaningfully numerically differentiable [*Kavetski et al.*, 2006a, Appendix A], hence ruling out objective functions such as those of explicit schemes in Figures 1 and 2. This explains our general preference for time stepping methods with good microscale properties. It is also worth noting that moderate numerical noise in the objective function can be handled by increasing the perturbation in the finite difference gradients used by Newton-type optimizers [*Dennis and Schnabel*, 1996], although this necessarily increases the

truncation error in the gradient approximation and hence has limited utility for exceedingly noisy functions.

8.4. Analysis of Multimodality

[68] While the benefits of global optimization are evident, it is worth noting that, somewhat ironically, local optimization may yield global insights not readily available from pure global optimization. In particular, the convergence behavior of the multi-start quasi-Newton method yields useful insights into the multimodal characteristics of the objective function that are not available in the SCE method [*Kavetski et al.*, 2006b; *Skahill and Doherty*, 2006; *Kavetski et al.*, 2007]. Such information on multimodality can be highly valuable. For example, consider the multimodal objective function in Figure 8 of *Thyer et al.* [1999], where the global mode of the SFB model lies on a parameter bound, indicating some model degeneracy. Yet the three local optima in the interior of the parameter space appear associated with much larger high-probability regions of the parameter space and hence may have larger total probability masses, and moreover, may well correspond to more meaningful model structures. Indeed, in cases where several local optima yield comparable model performance, it may be the observation and measurement errors in the calibration data that ultimately determine which optimum becomes dominant (global). Consequently, the multi-start quasi-Newton optimizer, which more readily diagnoses and reports secondary optima, may be more informative than a purely global optimizer. Importantly, in our numerical trials, the computational efficiency of each quasi-Newton sequence often allowed a complete analysis of the objective function at a comparable cost as a single global optimization using the SCE search! These insights could be used to improve the model structure and to derive more identifiable models.

[69] In this study, Figure 6 suggests that the objective functions of FUSE-070 and FUSE-060 contain secondary local optima even when near-exact solution of the governing equations is obtained using adaptive substepping with a moderate tolerance (τ_R = 1% and τ_A = 0.01 mm). This confirms the earlier caution by *Kavetski et al.* [2006a] that numerical artifacts may explain most, but not necessarily all, multimodality encountered in hydrological calibration. In particular, sum-of-squares surfaces of nonlinear models cannot be guaranteed to be unimodal except in very restrictive cases [*Demidenko*, 2000] (see also discussion by *Bates and Watts* [1988]), which are unlikely to be fulfilled by most practical hydrological models. Hence, we caution the reader that robust numerical approximations, indeed, even exact solutions, of most nonlinear hydrological models cannot guarantee unimodality of their objective functions. However, they will (virtually) guarantee absence of *spurious* multimodality due to numerical artifacts and bring a host of other benefits as discussed elsewhere in this paper and its companion.

[70] To more broadly evaluate the impact of the time stepping scheme on spurious multimodality, Figure 8 shows the distribution of local optima for different time stepping solutions of all 6 FUSE models applied to each of the 12 MOPEX basins. Figure 8 confirms that models implemented using the fixed-step explicit Euler and explicit Heun schemes have more local optima at lower objective function values than the models implemented with the fixed-step implicit Euler and adaptive explicit Heun solutions. This result is most pronounced for models FUSE-060, FUSE-536, and FUSE-550, which, not coincidentally, also had some of the roughest objective functions when solved using uncontrolled explicit approximations. Since most of this multimodality disappears when more robust time stepping solutions are implemented, this is indeed spurious multimodality arising from numerical artifacts of erratic time stepping schemes.

8.5. Verifying Multimodality: Limitations and Pitfalls

[71] We stress that reliable detection of multimodality in N_d-dimensional functions $\Phi(\boldsymbol{\theta})$ can be tricky even for low N_d. For example, if the function contains multiple optima of the same value (such multiple modes may be jointed or disjointed), convergence of multiple optimization sequences to the same objective function value does not guarantee that the same optimum is found. This can be diagnosed by inspecting the parameter sets at which the individual sequences have terminated. On the other hand, termination at distinct parameter sets may itself falsely suggest multimodality if the termination was due to a failure of the optimizer. Hence, while such failures should be rare for Newton-type optimization of smooth functions, it is best to force tight convergence criteria (to avoid false convergence) and directly check the optimality criteria upon termination: near-zero gradient $\partial\Phi/\partial\boldsymbol{\theta}$ and positive-definite Hessian matrix $\partial^2\Phi/\partial\boldsymbol{\theta}^2$ (for minimization in the interior of the parameter space). Analogous criteria hold for optima on parameter constraints and when maximizing rather than minimizing [*Gill et al.*, 1981].

[72] In addition, as illustrated in Figure 9, visual assessments of multimodality using lower-dimensional slices (projections) and marginals must also be approached with considerable caution, since they can easily lead to misleading conclusions even for comparatively simple-shaped objective functions (parameter distributions) [see also *Kavetski et al.*, 2006b, p. 192]. For example, depending on the direction of the slice and the geometry of the objective function, cuts through the objective function can fail to detect multimodality or, conversely, falsely indicate it. Figure 9 also depicts similar problems when trying to establish the multimodality of a joint probability distribution by examining its marginal densities, e.g., such as those constructed from MCMC samples.

8.6. Microscale Continuity

[73] In addition to macroscale features, parameter optimization is also affected by the microscale characteristics of the objective function [e.g., *Gill et al.*, 1981; *Kavetski et al.*, 2006a]. In particular, microscale discontinuities introduced by adaptive substepping (e.g., Figure 2) degrade the representativeness of the gradient and Hessian of the objective function in determining an efficient search direction, which ultimately delays or even impedes the optimization. The variable number of Newton-Raphson iterations in implicit schemes also introduces nonsmoothness, but this can be reduced by enforcing a tight Newton-Raphson convergence tolerance [*Kavetski et al.*, 2006a].

[74] Interestingly, Figure 8 illustrates that the microscale roughness caused by adaptive time stepping is not necessarily detrimental: The distributions of Nash-Sutcliffe optima obtained using multiple quasi-Newton sequences applied to the nonsmooth adaptive solutions are very similar to those obtained for the smooth fixed-step implicit Euler

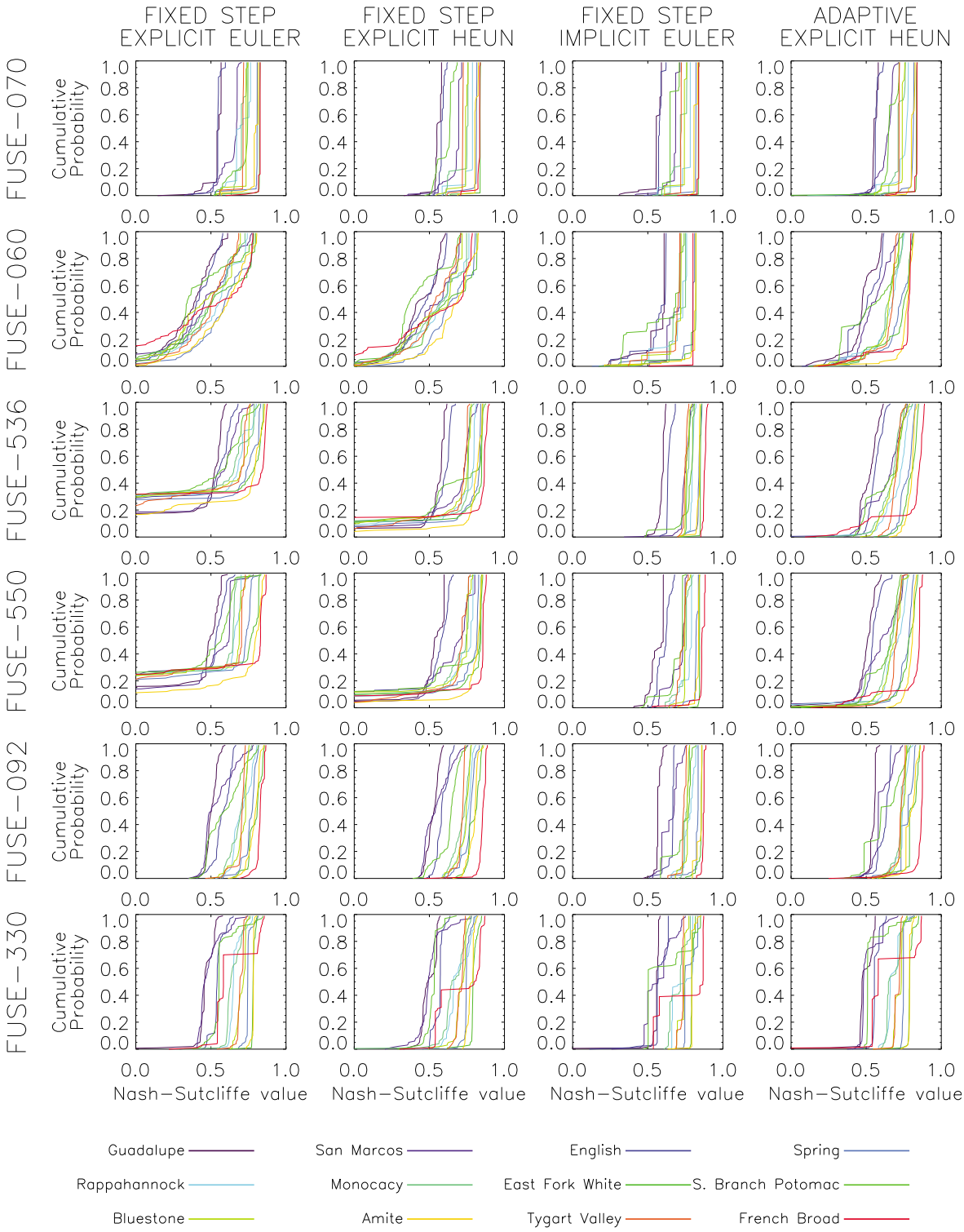

Figure 8. Distribution of local optima identified using the multi-start quasi-Newton method for different time stepping implementations of the 6 FUSE models, for all 12 MOPEX basins. Models implemented using fixed-step explicit Euler and Heun methods have considerably more local optima in their objective functions than those solved using the fixed-step implicit Euler and adaptive explicit Heun schemes (less steep cdf's of local optima). Note that, regardless of numerical artifacts, the multimodality structure of the objective function depends on the hydrological model and the calibration data.

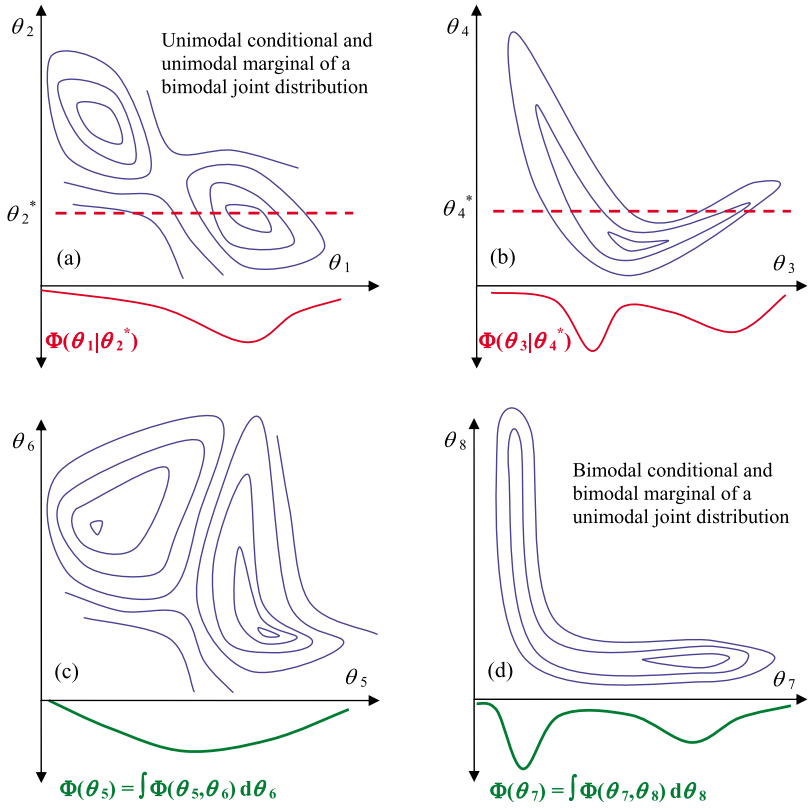

Figure 9. Potential pitfalls in the visual assessment of the multimodality of an objective function (joint parameter distribution) $\Phi(\boldsymbol{\theta})$ (blue contours) using lower-dimensional slices (dashed red lines and red profiles) and marginals (green profiles), illustrated using four hypothetical two-parameter models. (a and c) Cross section and the marginal fail to reflect the multimodality of the joint distribution. Multimodality suggested by the cross section and the marginal is not present in the joint distribution. In addition, (b) near-optimal values of parameter θ_3 correspond to a trough in the conditional cross-section, and (d) the tail regions of parameter θ_7 in the joint posterior dominate near-optimal values when the marginal is considered. These effects are not "artifacts" but can be counterintuitive and confounding to a modeler; Figures 9a, 9b, and 9d are likely to be quite common in practice. Although the presentations are based on 1-D projections and marginals along the parameter axes of two-parameter models, the same problems will arise for 2-D and other multivariate slices and marginals of higher-dimensional distributions, nonorthogonal cross sections and marginals, etc. In general, an N_d-dimensional function cannot be uniquely summarized solely by analysis of its lower-dimensional subspace or marginals. In the context of parameter optimization, the optima suggested by a numerical optimizer, such as the multi-start quasi-Newton algorithm, may therefore need additional verification (e.g., by inspecting the location, gradient, and Hessian of individual candidate optima). Such verification is much easier for smooth numerically differentiable functions.

approximation. This is consistent with the studies by *Kavetski et al.* [2003b, 2007] and *Skahill and Doherty* [2006], where Newton-type schemes with finite difference gradients successfully optimized models that, strictly speaking, are not smooth. However, this is likely case-specific and depends on the extent and type of non-smoothness.

[75] The multibasin multimodel analysis, summarized in Figure 8, indicates cases where optimization of models solved using the (nonsmooth) adaptive explicit Heun method terminates at notably lower objective function values than those based on the (smooth) fixed-step implicit Euler scheme. Again, this was particularly pronounced in FUSE-536 and FUSE-550, where runoff is often dominated by interflow [*Clark and Kavetski*, 2010] and the gradients contained appreciable numerical noise (Figure 2). This finding suggests that the fixed-step implicit Euler approximation with a tight Newton-Raphson tolerance, while less accurate in a strict numerical sense due to time discretization errors, may nonetheless be favored in the model optimization context due to its smoothness. We also refer the reader to *Gill et al.* [1981] for an authoritative discussion of adjusting model implementation to avoid harmful numerically induced nonsmoothness.

8.7. Broader Implications for Hydrology

[76] The benefits of multi-start Newton-type optimization suggested in this paper have several significant methodological implications for hydrological calibration. Recall that *Duan et al.* [1992] introduced the SCE method to address problems typical in the calibration of hydrological models. These problems include (1) multiple convergence regions (multiple large-scale optima or regions of attraction), (2) many

small "pits" in each region of attraction, (3) rough response surface with discontinuous derivatives, (4) poor and varying sensitivity of the response surface in the region of the optimum and nonlinear parameter interactions, and (5) nonconvex response surfaces with long curved ridges [*Duan et al.*, 1992]. These problems motivated the abandonment of "fast" derivative-based optimization in favor of "slow" global evolutionary searches.

[77] This study provides stronger empirical evidence supporting the claims of *Kavetski and Kuczera* [2007]: Many of the calibration problems outlined by *Duan et al.* [1992] are caused by poor numerical implementation of hydrological models and can be largely eliminated by using robust numerical time stepping schemes and by using modern quasi-Newton optimizers with trust-region (or linesearch) safeguards to ensure steady progress toward the (nearest) optimum. The feasibility and competitiveness of Newton-type optimization for conceptual hydrological models is also indicated by recent successful applications of quasi-Newton methods to simple versions of TOPMODEL and VIC [*Kavetski et al.*, 2003b, 2006a] and to the Six Parameter (SIXPAR) model (but not the Soil and Water Assessment Tool (SWAT2000)) [*Shoemaker et al.*, 2007; *Tolson and Shoemaker*, 2007], as well by successful Gauss-Newton optimization of the Hydrologic Simulation Program Fortran watershed model [*Skahill and Doherty*, 2006]. As illustrated and discussed above, the ability to apply powerful gradient-based Newton-type methods opens new avenues to drastically simplify hydrological model calibration while actually providing additional insights into the parameter distributions.

8.8. Further Improvements

[78] In this work the objective function gradient needed for the quasi-Newton method is approximated using finite differences, one-sided initially and switching to more accurate (for a smooth function!) central differences near the optimum. Since these finite differences require, respectively, N_d and 2 N_d objective function calls, the majority of the computational cost of the quasi-Newton scheme in this study is actually gradient approximation! It follows that substantial increases in efficiency of the quasi-Newton optimizer are possible if analytical derivatives are used: (1) it obviates finite differencing objective function calls and (2) it avoids finite difference gradient errors and hence allows even faster progress and more reliable termination of each search sequence.

[79] On the basis of pilot studies with VIC-type conceptual models [*Kavetski et al.*, 2007], gains in efficiency by factors of N_d or more are generally achieved when finite difference gradients are replaced by exact evaluation. However, while analytical differentiation is relatively straightforward to derive and implement for a fixed model structure [e.g., *Gupta and Sorooshian*, 1985], it is tedious and considerably more programmatically difficult to incorporate into a flexible model software such as FUSE. It is therefore left for future work.

[80] A combination of global searches with gradient-based optimization may be beneficial. However, Figure 7 suggests that standard approaches that apply gradient-based Newton-type optimizers to seeds already evolved using global optimization [e.g., *Tolson and Shoemaker*, 2007] may not be advantageous, indeed the opposite approach of mopping up minor improvements using gradient-free methods, such as the simplex algorithm, may be warranted [*Gill et al.*, 1981], especially for micro-noisy hydrological models such as those implemented using adaptive time stepping.

[81] In addition, the Hessian approximation constructed as part of the quasi-Newton optimization can be used to initialize (or update) the covariance matrix of subsequent MCMC sampling (section 5.4.2) (as discussed, e.g., in the study by *Kavetski et al.* [2006b], the negative inverse Hessian provides an estimate of the covariance for near-Gaussian probability distributions). Furthermore, the computation of the objective function gradient elements can be carried out in parallel. These enhancements are left for the future.

[82] The hybrid strategies above are in line with recent developments in global optimization, which seek to combine multiple methods such as genetic algorithms, simulated annealing, swarms, and ant colonies, etc., for handling complex objective functions [e.g., *Vrugt and Robinson*, 2007]. However, an overriding message of this paper, exemplified in Figure 1, is that addressing a root cause of the problem, namely numerical artifacts, significantly simplifies the hydrological calibration challenge and, in some cases, may reduce the need for more complicated calibration procedures.

[83] In special cases, we suggest going even further and actually modifying the governing equations themselves to eliminate or replace components known to lead to problematic objective functions. For example, model thresholds can often be smoothed with minimal, if any, loss of model performance [e.g., *Kavetski et al.*, 2006a; *Kavetski and Kuczera*, 2007]. Indeed, the evaporation-storage relations in FUSE are smoothed (Table 2 of the companion paper). Note that such genuine model modifications require a sound judgment by the modeler and must be approached with care.

9. Impacts on Inference and Prediction

9.1. Parameter Inference

[84] To examine the impact of the time stepping scheme on uncertainty estimation, Figure 10 shows the bivariate marginal parameter distribution (estimated using MCMC sampling) for all 6 FUSE models applied to the Mahurangi basin. For space limitations, only the parameters common to all six FUSE models are displayed (others exhibit a similar behavior).

[85] Figure 10 indicates that parameter inference is extremely sensitive to the numerical time stepping scheme used to solve the governing equations. The parameter distributions of models implemented using fixed-step explicit methods were almost always very distant from those corresponding to the near-exact solution (here, indistinguishable from the adaptive explicit Heun solution). In many cases, the differences are staggering, e.g., the maximum unsaturated storage $S_{1,max}$ was estimated at around 50 ± 10 mm in FUSE-060 implemented using the fixed-step explicit Euler approximation, whereas for the near-exact solution of the same model equations, it was close to 500 ± 50 mm.

[86] Importantly, even generally fidelious methods, such as the fixed-step implicit Euler scheme, can have very different parameter distributions from the same models solved near-exactly (e.g., for FUSE-550). However, in contrast to uncontrolled explicit time stepping, the parameter estimates obtained for models solved using the fixed-step implicit Euler approximation were generally close to the values obtained when adaptive time stepping was used (e.g., for FUSE-060).

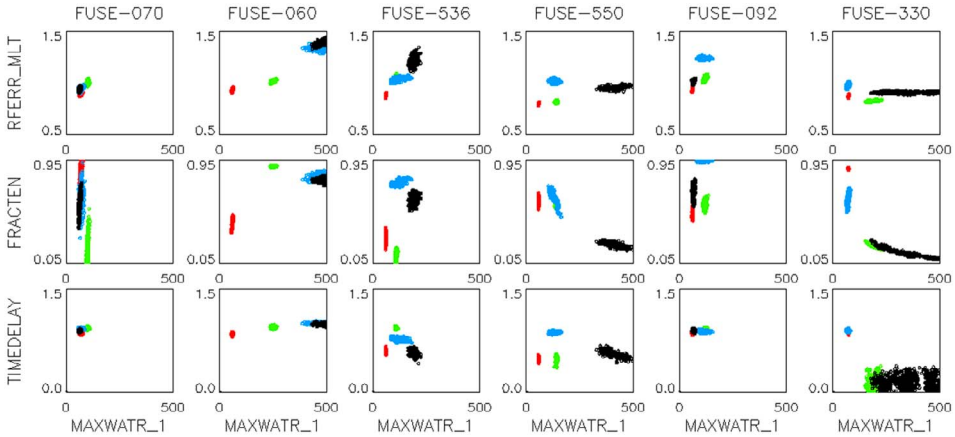

Figure 10. Bivariate marginal parameter distributions for common parameters of the six FUSE models applied to the Mahurangi basin. The colors denote the fixed-step explicit Euler (red), fixed-step explicit Heun (green), fixed-step implicit Euler (blue), and adaptive explicit Heun (black) schemes. The impact of the time stepping scheme on the calibrated parameter values and uncertainty estimates is striking in all models but especially in FUSE-550.

9.2. Comment on Alternative Sampling Techniques

[87] We stress that similar results would have been obtained using alternative analysis techniques, such as Monte Carlo Importance sampling [*Kuczera and Parent*, 1998], Gibbs sampling [*Reichert et al.*, 2002], Shuffled Complex Evolution (Metropolis) (SCEM) and Differential Evolution Adaptive Metropolis (DREAM) algorithms [*Vrugt et al.*, 2003, 2009], and others. Indeed, given the severe deformations of the models' objective functions shown in Figure 1, any convergent sampling method will necessarily produce parameter estimates and associated distributions that reflect these distortions. There is simply no way for a sampling scheme, regardless of its sophistication, to detect, let alone correct, numerical errors of the underlying hydrological model implementation.

[88] Moreover, such deformations may delay, or even prevent, convergence of Monte Carlo samplers, yielding results that, in addition to being corrupted by truncation errors of the time stepping approximation, are also subject to sampling artifacts due to incomplete convergence of the sampling algorithm. Markov chain sampling could be particularly vulnerable to getting trapped on macroscale optima [e.g., *Vrugt et al.*, 2009; *Schoups et al.*, 2010], but importance samplers will also converge slower for highly irregular probability distributions, and importantly, these irregularities would considerably complicate any adaption of the importance function. Finally, while increasingly powerful MCMC methods are being developed to handle geometrically complex target distributions [e.g., *Vrugt et al.*, 2009], our current opinion is that robust solutions for reliably sampling from multimodal distributions with a priori unknown modes separated by vast low-probability spaces are yet to be developed. Searching for such high-probability narrow-support "islands" is an extremely challenging proposition, cursed by the sheer volume of high-dimensional spaces. Moreover, practical diagnostics will often falsely indicate convergence if a high-probability mass region was missed by the sampler. It is likely that successful solution strategies may require some restrictions on the nature of the target probability distribution, independent insights, and a combination of multiple numerical submethods including optimization and adaptive MCMC. Hence, while looking forward to such developments, we emphatically recommend removing at least spurious deformations of the model's objective function. Using robust time stepping schemes is one such essential "artifact-prevention" strategy.

9.3. Model Predictions and Internal Dynamics

[89] Ultimately, the operational objective of conceptual hydrological models is the prediction of streamflow and, in some cases, insights into the internal catchment dynamics such as storage (though the interpretation of internal states in conceptual models is often unclear and requires considerable caution, see *Duan et al.* [2006]). In this section, we inspect the impact of time stepping on the streamflow predictions, as well as on the estimates of internal model storages.

[90] As determined in section 8.2, the peak Nash-Sutcliffe performance of all models was comparable across all time stepping schemes, yet section 9.1 indicated massive differences in the corresponding parameter inference. This leads to the question: Is it possible that the differences in the inferred parameters simply reflect strong interdependencies and compensations, and the optimized models have similar response and internal dynamics regardless of the time stepping implementation?

[91] To address this question, Figure 11 compares the optimized streamflow time series and corresponding upper zone storages predicted using four different time stepping schemes, for each of the six FUSE models calibrated to the Mahurangi basin. It shows that, for any given FUSE model, there are only very minor differences between the optimized streamflow predicted using different time stepping schemes (Figure 11, left). This is consistent with previous findings (e.g., Figure 6): There are always some parameter sets that provide a good fit to the observations, even for time stepping scheme with poor overall fidelity, and the optimization method, whether Newton-type or SCE, is able to reliably find these parameter sets.

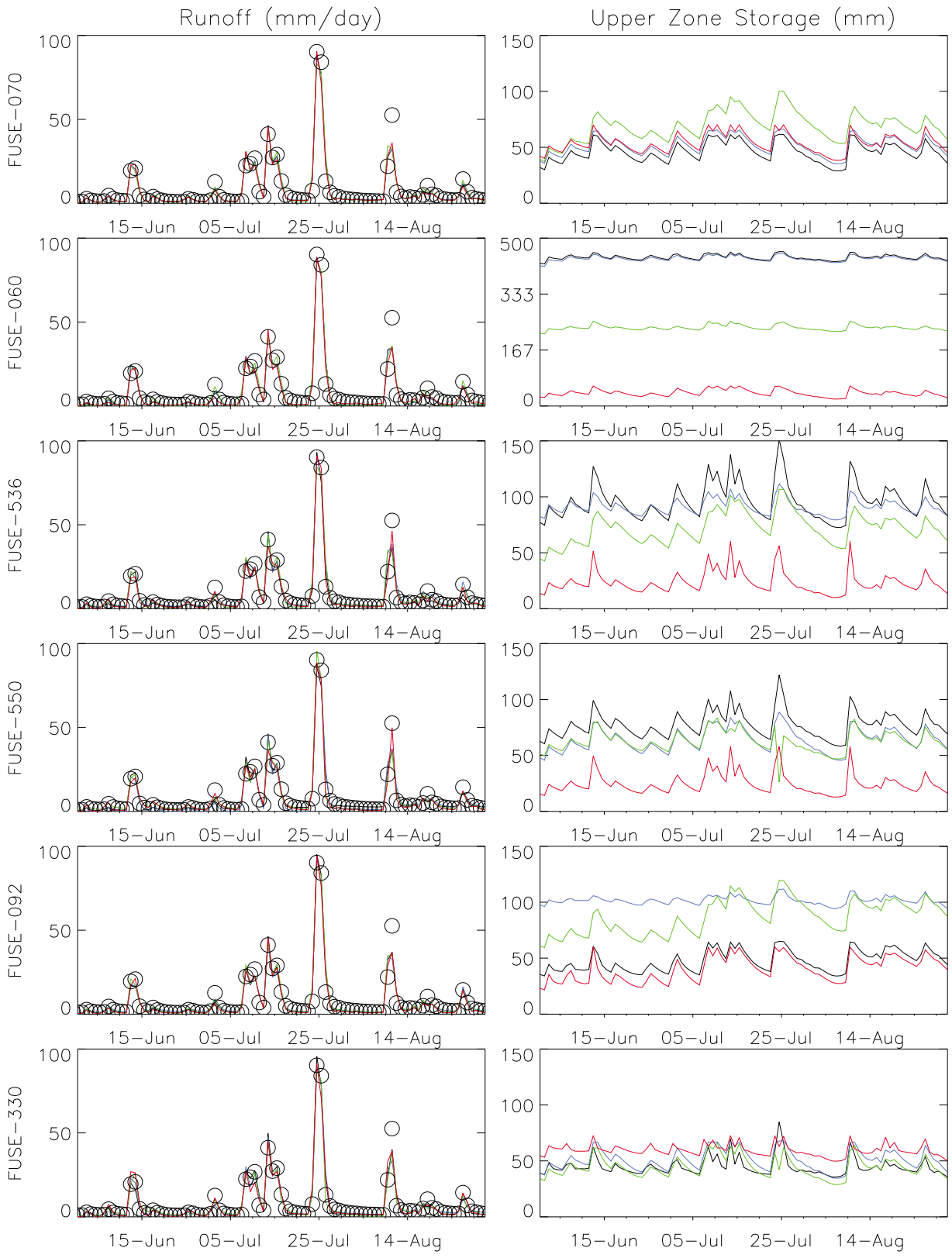

Figure 11. Simulations of (left) runoff and (right) upper zone storage in the Mahurangi basin, for each of the six FUSE models, using daily rainfall forcing and near-optimal parameter sets. The colors denote the fixed-step explicit Euler (red), fixed-step explicit Heun (green), fixed-step implicit Euler (blue), and adaptive explicit Heun (black) schemes, whereas the circles denote observations. The optimized runoff time series are indistinguishable from the fixed-step explicit estimates due to parameter compensations. Despite near-identical streamflow predictions, note the considerable differences in corresponding storage levels, depending on the time stepping scheme (right plot).

[92] However, even when the predicted streamflows are near-identical, simulations of upper zone storage depend markedly on the time stepping schemes and associated optimal parameters (Figure 11, right). Characteristically, this is especially pronounced for fixed-step explicit schemes in virtually all FUSE models considered in this study. While determining the precise reasons for these discrepancies requires detailed and case-specific truncation error analysis beyond the scope of this study, we make the following observations: (1) since truncation errors depend on the model parameters, optimization of the latter may result in truncation errors canceling model structural errors for nonbehavioral parameter sets, and (2) matching the total streamflow does not imply that the individual constituents, e.g., surface runoff, interflow, and baseflow, are preserved.

[93] Interestingly, simulations of upper zone storage are quite similar in the fixed-step implicit Euler and (near-exact) adaptive explicit Heun schemes, even in cases where the parameter distributions differ markedly. For example, models FUSE-536, FUSE-550, and FUSE-330 have very different parameter distributions when implemented using the fixed-step implicit Euler versus the adaptive explicit Heun schemes, yet the simulations of upper zone storage are very similar (here, suggesting compensatory effects between the maximum water content and fractional tension storage parameters). However, such compensatory behavior is not universal; for example, the upper zone storage predicted by FUSE-092 solved using the fixed-step implicit Euler scheme is quite different from that obtained when the same equations are solved with adaptive error control. These findings reinforce the fact that unconditional stability does not guarantee strict numerical accuracy.

9.4. Validation

[94] Figure 11 suggests that, regardless of the inferred parameter values and internal dynamics, the fitted (optimized) model streamflows are remarkably similar even when different time stepping schemes, including the fixed-step explicit Euler approximation, are used. Let us put aside, just for a moment, the (likely) possibility that the numerical errors in the fixed-step explicit Euler scheme may cause "behavioral" parameters to be classified as "nonbehavioral" and vice versa. Let us also ignore their confounding effects on the interpretation of internal process dynamics. Let us ask, in the spirit of "black-box" modeling, the following question: If the fitted streamflows produced using the fixed-step explicit schemes are just as accurate, is it really necessary to implement numerically reliable time stepping methods? Indeed, could it be that optimization has tamed the capricious fixed-step explicit beast, making it just as well behaved as unconditionally stable or adaptive solutions?

[95] To address this question, we first recognize that indeed streamflow predictions are the primary quantity of interest in many (though not all) operational applications of hydrological models. However, while the performance during the calibration period is of clear interest to a practitioner, their key priority is usually performance under different forcing conditions. Indeed, notwithstanding debates regarding whether models can be validated [*Konikow and Bredehoeft*, 1992], at least some kind of split-sample testing is widely recognized as necessary, though not necessarily sufficient, if a model is to have predictive credibility [*Klemes*, 1986; *Kuczera and Franks*, 2002].

[96] Thus, to address the pragmatic black-box modeler's question, we evaluated ("validated") the performance of different time stepping schemes over a series of yearly independent "validation" periods that were not used in the calibration (section 4.2). We then evaluated the "predictive fidelity" of the fixed-step explicit and implicit Euler schemes, using the adaptive explicit Heun as a surrogate near-exact solution of the governing equations (its adequacy was verified in several tests using tighter truncation error tolerances). The predictive fidelity is computed using the same equation as the fidelity measure defined in the companion paper,

$$\varphi_{XX}^{(abs)} = \| \tilde{\mathbf{y}} - \mathbf{y}_{XX} \| - \| \tilde{\mathbf{y}} - \mathbf{y}_{exact} \|, \qquad (11)$$

where $\mathbf{y}_{exact}$ and $\mathbf{y}_{XX}$ are the exact and numerical solutions of the model's governing equations, respectively; and $\tilde{\mathbf{y}}$ is the observed data. While in the companion paper $\tilde{\mathbf{y}}$ represented the calibration data, here we apply equation (11) over the validation period. This gauges the ability of the numerical scheme to reproduce the exact solution in the context of fitting validation data, which is evidently a more stringent check than reproducing the calibration data. It is also a more pertinent check: It tests the hydrological model for its intended application, which is predicting unknown streamflow.

[97] Figure 12 compares the predictive fidelity of the fixed-step explicit and implicit Euler approximations versus the predictive fidelity of the near-exact solution of the model equations. The analysis includes all 19 years in the validation period, with all six FUSE models applied to the French Broad and Guadalupe River basins (which are, respectively, the wettest and driest of the 12 MOPEX basins). Two observations are immediately apparent.

[98] 1. There is a markedly larger difference in validation performance between the fixed-step explicit Euler and adaptive explicit Heun solutions than between the fixed-step implicit Euler and adaptive explicit Heun solutions (compare Figure 12 (left) and Figure 12 (right)).

[99] 2. In the French Broad River, the fixed-step explicit Euler approximation suffers, on average, from an unequivocally larger degradation in performance during the validation period than the fixed-step implicit Euler solution (the top left plot of Figure 12). The results for the Guadalupe are comparable, although the differences are smaller, and in some cases, there is a fortuitous improvement.

[100] To further illustrate the loss of validation performance, Figure 13 depicts representative streamflow time series from the validation experiments. Note that while even the near-exact ODE solution is unable to fit the third storm event, the fixed-step explicit Euler approximation is markedly worse than the fixed-step implicit Euler solution. Indeed, it introduces 5 mm/d errors versus 0.1 mm/d of additional error. The magnitude, both absolute and relative, of these additional errors is reminiscent, not coincidentally, to the fidelity evaluations undertaken in the companion paper.

[101] The markedly worse degradation of the fixed-step explicit Euler approximation vis-à-vis its implicit counterpart is not surprising, given the general numerical fragility of uncontrolled explicit schemes. In any numerical ODE method, the truncation errors, which comprise high-order derivatives of the solution, depend on the forcing regime and on the current model states. However, in methods that are only

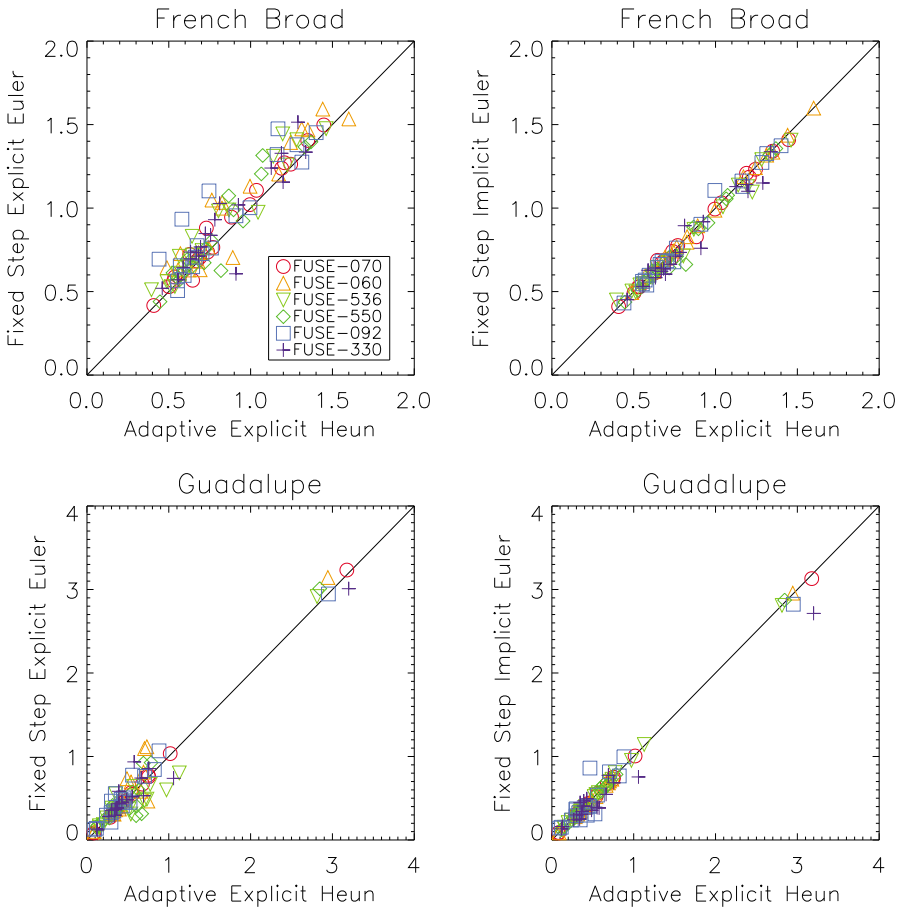

Figure 12. Impact of time stepping scheme on model validation and predictive performance: (left) predictive fidelity of the fixed-step explicit Euler and (right) fixed-step implicit Euler approximations versus the predictive fidelity of the near-exact FUSE solution. Results are shown for all 19 individual year-long validation periods (1961–1979) and each of the six FUSE models, applied to the (top) French Broad and (bottom) Guadalupe. Fidelity is computed using equation (11). The implicit Euler scheme has a noticeably higher predictive fidelity than the explicit Euler scheme, especially in the French Broad basin, and in general, faithfully approximates the underlying model equations. Conversely, uncontrolled explicit time stepping suffers larger losses of predictive performance under different forcing conditions encountered in operational use and is clearly unsafe.

conditionally stable, such as the explicit Euler scheme, truncation errors depend much more sensitively on these factors and, moreover, are readily amplified if the scheme becomes even marginally unstable. This can easily happen for new forcing data regimes that may be encountered in predictive operational use.

9.5. Implications for Hydrological Modeling

[102] While the similarity of predicted streamflows creates an illusion of adequacy of the fixed-step explicit Euler scheme, several comments can be made.

[103] 1. If fitting the observed streamflow is the sole objective of the model application, even fixed-step explicit schemes appear to provide satisfactory optimal results, although finding optimal parameters is much harder and more expensive because the objective function is poorly behaved (Figure 1). What effectively happens is that numerical truncation error is added to otherwise poor simulations to give better results. This also destroys the performance for parameter sets giving good fits to the data when the model is solved accurately (e.g., see Figure 1 in the study by *Clark and Kavetski* [2010]).

[104] 2. If consistency of inferred parameters, and/or inference of internal dynamics, is of interest, and especially if interpretation or comparative analysis of these inferences is desired, numerical errors arising in unreliable time stepping schemes will preclude the investigator from reaching meaningful conclusions. We suspect that numerical artifacts have played a major role in confounding model evaluation and comparison experiments and, if left unattended, will continue to impede meaningful progress in conceptual hydrological modeling and engineering.

[105] 3. If predictive credibility of the model is important (we struggle to imagine when this would not be the case), fixed-step explicit schemes tend to perform markedly worse in validation than fixed-step implicit or adaptive explicit solutions. We would be particularly wary of attempting to extrapolate a model implemented using uncontrolled explicit time stepping, as it is liable to behave even more erratically and unpredictably. Conversely, unconditionally stable schemes

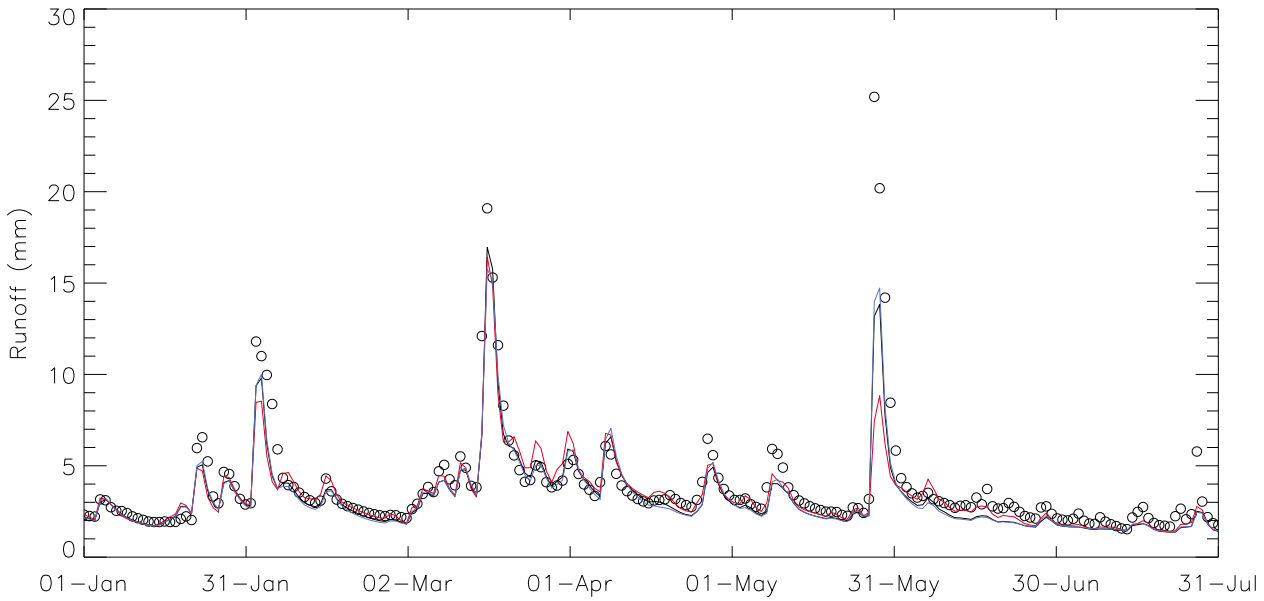

Figure 13. Representative validation streamflow time series for FUSE-092 applied over the 1973 period in the French Broad River basin. Predictions obtained using the fixed-step explicit Euler (red), fixed-step implicit Euler (blue), and adaptive explicit Heun (black) schemes are compared to observed flows (circles). The fixed-step implicit approximation is very similar to the near-exact adaptive solution, and both have markedly better predictive performance for the large storm event at the end of May than the fixed-step explicit Euler approximation. Particularly strong deterioration of predictive performance for the fixed-step explicit Euler scheme is common, as seen in Figure 11.

(implicit Euler), or schemes based on direct error control (e.g., adaptive explicit Heun), are much less fragile in this respect, and their discrepancy with the observations generally reflects genuine structural errors of the model's governing equations and/or errors in the model forcing data.

9.6. Generality of Results

[106] To evaluate the generality of the results presented in Figures 10 and 11, Figure 14 presents bivariate marginal distributions for the parameters of the FUSE-070 model applied in all 12 MOPEX basins. In general, the fixed-step explicit Euler and explicit Heun methods yield very different parameter distributions than the fixed-step implicit Euler and adaptive explicit Heun methods. This is expected given the macroscale distortions in the objective function surfaces occurring in fixed-step explicit implementations. The parameter distributions for the fixed-step implicit Euler and adaptive explicit Heun methods are generally quite similar (Figure 14), but in some cases, there are notable differences, especially in the San Marcos (SAN), English (ENG), Rappahannock (RAP), and South Branch Potomac (POT) basins. The multibasin analysis confirms that the choice of time stepping scheme can considerably affect parameter inference, especially for unreliable fixed-step explicit methods, but, notably, also even for unconditionally stable implicit schemes when error control is not implemented.

10. Practical Selection of Time Stepping Schemes for Conceptual Hydrological Models

10.1. General Considerations

[107] Practical assessment of adaptive time stepping schemes depends on trade-offs between numerical reliability and computational cost. While adaptive time stepping methods may match the exact solution if the truncation tolerance is sufficiently tight, the resulting computational cost can be prohibitive, especially for low-order methods. For example, the near-exact adaptive Heun solution obtained by setting the truncation tolerance close to machine precision required thousands of flux evaluations per time step [*Clark and Kavetski*, 2010]. This is infeasible and, moreover, unnecessary in most practical hydrological applications. Slacker tolerances, e.g., constraining truncation errors below 1%, drastically cut the cost while maintaining numerical errors well below those likely to arise due to model structure and uncertainty in forcing data such as rainfall. We stress that the opposite is true in uncontrolled explicit methods, where numerical errors frequently dwarf structural and forcing errors, even under common hydrological conditions [*Clark and Kavetski*, 2010].

10.2. Fixed-Step Implicit Euler Versus Adaptive Explicit Heun Methods

[108] This paper indicates that the fixed-step implicit Euler method may represent a reasonable practical alternative to adaptive time stepping methods. Yet, given the clear accuracy and efficiency advantages of the adaptive explicit Heun method with a moderate truncation tolerance (Figure 5), we revisit the earlier question: Is there a place for fixed-step methods in conceptual hydrological models? Table 1 compares the adaptive explicit Heun scheme with several fixed-step methods.

10.2.1. Accuracy and Fidelity

[109] Like the adaptive explicit Heun method, the fixed-step implicit Euler method provides a faithful approximation of the exact solution, in particular, free of macroscale dis-

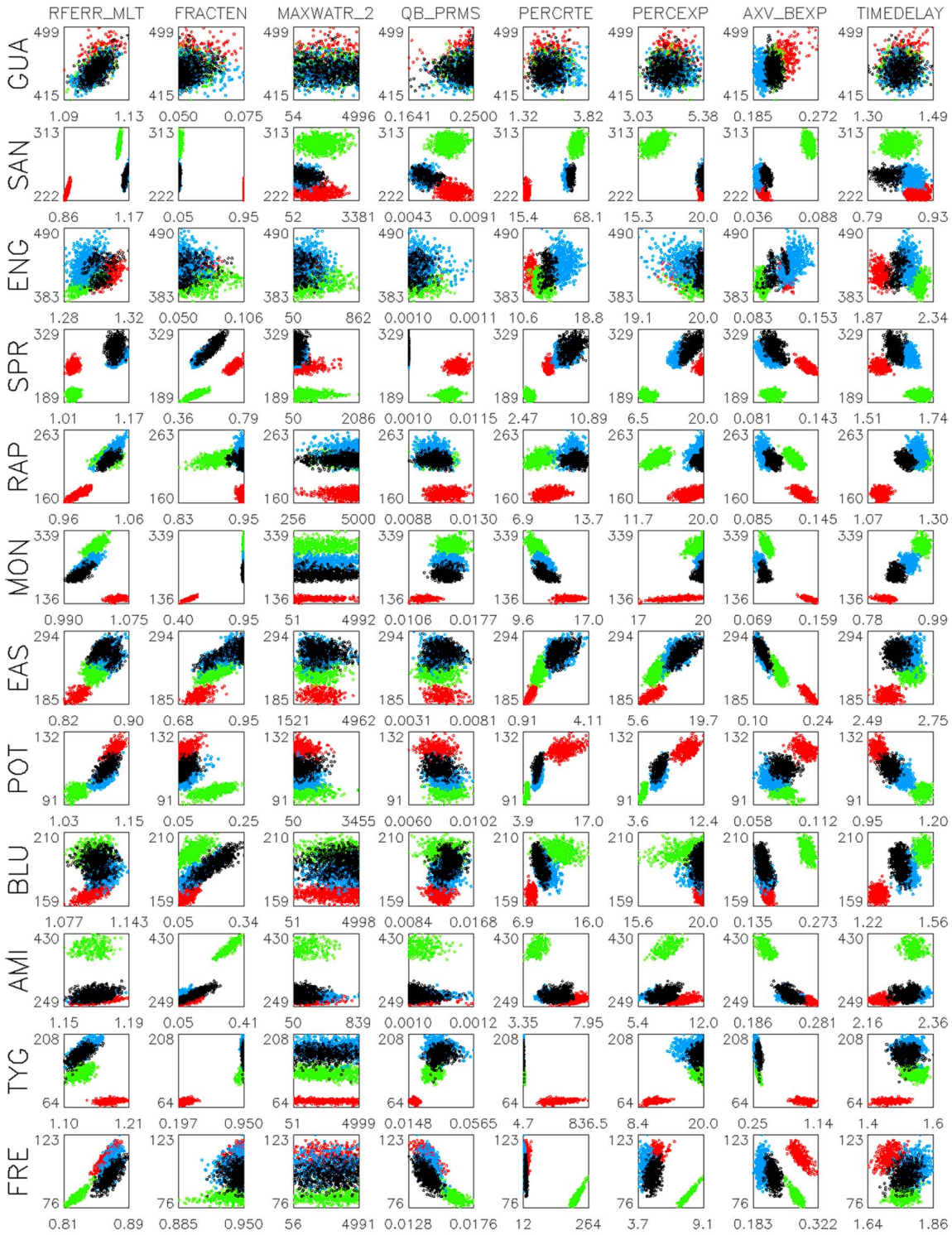

Figure 14. Bivariate marginal distributions of FUSE-070 parameters (x axis) versus $S_{1,\max}$ (y axis), for all MOPEX basins, estimated using MCMC sampling. The colors denote the fixed-step explicit Euler (red), fixed-step explicit Heun (green), fixed-step implicit Euler (blue), and adaptive explicit Heun (black) schemes. The impact of time stepping errors is evident, with the parameter distributions depending unpredictably on the time stepping scheme. MCMC convergence artifacts are also likely for geometrically complex distributions such as those arising as a result of fixed-step explicit approximations. The distributions arising from fixed-step explicit schemes are particularly inconsistent with the near-exact adaptive Heun solution, with the fixed-step implicit Euler approximation also exhibiting some discrepancies.

Table 1. Qualitative Comparison of the Adaptive Explicit Heun Solution to Fixed-Step Approximations

Methodological Considerations	Fixed-Step Explicit Euler	Fixed-Step Explicit Heun	Fixed-Step Implicit Euler	Fixed-Step Implicit Heun	Fixed-Step Semi-implicit	Adaptive Explicit Heun[a]
Free of macroscale distortions	✗	✗	✓	✓	✗	✓
Free of microscale discontinuities	✗	✗	✓[b]	✓[b]	✗[c]	✗[d]
Automatically satisfies solution constraints	✗	✗	✓	✗	✗	✗[d]
Average flux calls per time step	1	2	~10[e]	~10[e]	$1 + N_s$[f]	3–5[g]

[a]For the problems in this study, adaptive implicit Euler and implicit Heun solutions have similar numerical properties but higher average computational cost [*Clark and Kavetski*, 2010, Figure 5].
[b]Provided a sufficiently tight Newton-Raphson tolerance is satisfied.
[c]Because of finite difference Jacobian approximation effects and externally-imposed bound constraints.
[d]Minor impact, directly controlled by the truncation error tolerance.
[e]Average of 6 FUSE models over 12 MOPEX basins but depends on Newton-Raphson tolerances (here constraining $\Delta\zeta^{(m)}$ and $\mathbf{r}(\zeta^{(m)})$ below 10^{-9} mm).
[f]Depends on the number of state variables N_s, due to finite difference Jacobian approximation.
[g]Average of the 6 models over the 12 MOPEX basins but depends on truncation error tolerance (here $\tau_R = 1\%$ and $\tau_A = 0.01$ mm).

tortions. In the application of six different model conceptualizations to 12 diverse basins, the fixed-step implicit Euler scheme incurred numerical errors generally below 1% of the total error, thus providing adequate accuracy for most hydrological applications [*Clark and Kavetski*, 2010]. Nonetheless, adaptive error control does provide a closer approximation to the exact ODE solution [*Clark and Kavetski*, 2010], and, importantly, its numerical accuracy is directly controlled by a user-specified error tolerance.

10.2.2. Smoothness

[110] Unlike the fixed-step implicit Euler method, the objective functions of models implemented using the adaptive explicit Heun scheme are contaminated by microscale discontinuities (e.g., Figure 3; see also Figure 2 and Appendix A in the study by *Kavetski et al.* [2006a]). While this numerical noise can degrade the efficiency and robustness of model optimization [e.g., *Gill et al.*, 1981], the empirical assessment in Figures 6 and 8 suggests this may not be fatal, at least when (1) the discontinuities are localized rather than contaminate the entire surface and (2) the gradient is approximated using suitable finite difference intervals.

10.2.3. Solution Constraints

[111] Another relevant practical issue is the handling of solution constraints [*Clark and Kavetski*, 2010]. Provided intermediate iterations are safeguarded, the implicit Euler solutions automatically satisfy all state constraints. On the other hand, virtually all other schemes, in particular, the adaptive explicit Heun method, require external modifications to handle solution constraints. This can be complicated and somewhat subjective in coupled multistate models. Fortunately, adaptive substepping ensures that the mass balance associated with these ad hoc flux corrections is small and is controlled by the truncation error tolerance. While the imposition of "external" constraints, such as nonnegativity of solution, can degrade the efficiency of the error control in identifying a suitable step size [*Shampine et al.*, 2005], this was not observed in our empirical tests.

10.2.4. Computational Cost

[112] Finally, the computational cost must be considered. In the 12-basin analysis reported in the companion paper, the fixed-step implicit Euler scheme was, on average, costlier than the adaptive explicit Heun method with moderate error tolerances (relative errors below 1% or absolute errors below 0.01 mm). While the implicit Euler scheme can be accelerated using linesearches and Jacobian refreshment strategies [*Clark and Kavetski*, 2010], the adaptive explicit Heun method is generally enviably fast, and this is an important practical consideration for many applications.

[113] On the other hand, the companion paper also shows that the (fixed-step) implicit Euler scheme is more computationally robust, i.e., has a less variable computational cost than the (adaptive) explicit Heun scheme. More generally, in all models considered in our experiments, we encountered parameter sets where the adaptive explicit Heun algorithm was much more expensive than the implicit Euler scheme. Since this may be indicative of "stiff" ODE behavior [*Clark and Kavetski*, 2010, Appendix A], dynamic switches from explicit to implicit approximations when stiffness is suspected could be beneficial in terms of constraining worse-case runtimes.

10.2.5. Overall Comparison

[114] Choosing between the adaptive explicit Heun solution and the fixed-step implicit Euler approximation is largely a question of judgment and perspective (see Table 1). From a classical numerical ODE analysis perspective, the adaptive scheme is clearly preferred, because it generally provides controllably closer agreement with the exact solution. However, the fixed-step implicit Euler has more favorable microscale smoothness characteristics (Figure 2), which is advantageous in gradient-based model parameter optimization. Moreover, its unconditional stability makes it very robust (as listed in section 3.3 of the companion paper, explaining its widespread use in engineering software, including "industry-standard" groundwater and multiphase flow simulators).

[115] Interestingly, the broad empirical assessments in this paper suggest that the lower numerical accuracy of the implicit Euler method and the microscale discontinuities in the adaptive explicit Heun method do not materially corrupt model analyses such as parameter sensitivity and optimization, although numerical errors of the implicit scheme can affect the inferred posterior distributions of the model parameters and internal model states under some scenarios. Given that in some cases one method can be vastly preferable to the other, the option of adaptive switching warrants further attention, in particular, switching to an adaptive implicit scheme when stiffness is suspected to be causing gross computational inefficiencies in adaptive explicit integration.

10.3. Benefits of Numerical Enhancements

[116] Hydrological practitioners may not fully appreciate how subtle modifications of numerical approximations can

have spectacular impacts on model performance. For example, the original FUSE software [*Clark et al.*, 2008] used an adaptive implicit Euler scheme with truncation error estimated by step-halving (with fixed error tolerances $\tau_R = 10^{-4}$ and $\tau_A = 10^{-4}$ mm) and solved the implicit system using the classic Newton-Raphson scheme that recomputes the Jacobian at each iteration. In retrospect, the original FUSE implementation was spectacularly inefficient: The step-halving method required 33% more steps than embedded error control, and the truncation error settings were unnecessarily stringent. Implementing the adaptive explicit Heun method (with moderate error tolerances), or the fixed-step implicit Euler method (with Jacobian refreshment strategies), decimated the average computational cost by several orders of magnitude. Given the utility of tool kits such as FUSE in elucidating structural errors in process representations, which necessarily requires massive number of models runs, as well as confidence that the results are not rendered meaningless by numerical artifacts, the case for expediently moving away from unreliable numerical implementations cannot be overstated.

11. Conclusions

[117] This paper compared the performance of different numerical time stepping algorithms in the context of model application and prediction, including sensitivity analysis, optimization, and statistical inference of model parameters, as well as prediction of streamflow and internal storage states. To ensure the broad pertinence and generality of our conclusions, we carried out a thorough assessment using 8 distinct time stepping schemes, 6 hydrological models of varying range of complexity, and 13 catchments with diverse physical and hydroclimatic attributes.

[118] Several important conclusions were reached as follows:

[119] 1. When a hydrological model is implemented using unreliable time stepping schemes, in particular, fixed-step explicit methods, its objective function will generally be severely deformed by numerical artifacts. The extensive analysis in this paper indicates that these deformations are not rare isolated instances but affect virtually any model structure, in any catchment, and under common hydroclimatic conditions. Such artifacts may well explain many reported difficulties that historically complicated parameter optimization and have led to entire calibration paradigms not reliant on well-behaved parameter distributions. While fidelious time stepping schemes cannot guarantee unimodality, they do produce "better-behaved" objective functions that are free of spurious local optima and are numerically differentiable.

[120] 2. Sensitivity analyses of models implemented using unreliable numerical schemes reflect the combined sensitivity of the model equations and the numerical approximation errors. In many common cases, the sensitivity estimate is actually measuring the sensitivity of truncation errors to the model parameter values. In contrast, sensitivity analyses of models implemented using the fixed-step implicit Euler and adaptive (explicit) methods describe, as intended, the sensitivities of the governing model equations and are free of numerical artifacts.

[121] 3. Numerical implementation using reliable time stepping schemes, such as unconditionally stable implicit or adaptive explicit algorithms, immediately enable fast and informative gradient-based Newton-type parameter optimization. When applied to accurately implemented models with minimal spurious multimodality and sufficient numerical smoothness, the quasi-Newton sequences converge very fast to near-optimal parameter values and, when implemented within a multi-start framework, yield useful insights into the multimodality structure of the model's parameter distributions. Conversely, while global optimization methods such as the SCE search are robust, they are generally slower than multi-start Newton-type optimization because (1) global convergence necessarily requires a broader and hence more expensive exploration of the search space and (2) they typically use much less information regarding the shape of the objective function behavior than, e.g., gradient-based trust-region Newton-type methods.

[122] 4. Erratic time stepping schemes lead to inconsistent inference of model parameters and internal states, even if the streamflow predictions appear reasonable. Somewhat disturbingly, parameters in models implemented using inaccurate time stepping schemes can compensate for numerical approximation errors. Given that numerical errors of fixed-step explicit schemes routinely dwarf structural errors even under common hydrological scenarios, the compensation of these errors by altered parameter values during calibration drastically affects the conclusions of parameter uncertainty assessment and prevents meaningful parameter interpretation and regionalization.

[123] 5. Even when parameter interactions allow getting the "right result for the wrong reasons," the estimated internal dynamics are markedly dependent on the time stepping scheme. More important, the model's performance in validation mode is markedly lower for the fixed-step explicit schemes, which is readily attributable to their conditional stability, uncontrolled accuracy, and consequent general numerical fragility. Hence, getting the right results for the wrong reasons is just as wrong in numerical computation as elsewhere in hydrology and science [e.g., *Kirchner*, 2006]. Indeed, such "right results" deteriorate for even moderate departures from the calibrated conditions—they are yet another gift from Pandora's box of numerical artifacts in hydrology.

[124] More generally, this paper demonstrates that time stepping schemes prevalent in current conceptual hydrological models are numerically unreliable. They easily lead to erroneous conclusions regarding model sensitivity and, more seriously, inconsistent inferences of model parameters, their distributions, and internal model dynamics. This obscures the comparison and interpretation of model parameters (for example, in regionalization studies) and complicates meaningful improvement of the model structure and parameterization. It also confounds the identification of both dominant and nondominant hydrological processes, and their behavior, in a given catchment.

[125] In our opinion, the difficulties in handling highly irregular multimodal objective functions and probability distributions make the elimination of objective function complexity a key design priority. This may include not only removing spurious numerical artifacts but also, whenever appropriate, modifications of the model governing equations to avoid poorly behaved components. These issues will become increasingly important in high-dimensional inference problems, such as those arising in spatially distributed physical models and in hierarchical Bayesian estimation,

where the reduction of the computational burden becomes a key practical consideration.

[126] The traditional prevalence of fixed-step explicit time stepping in conceptual hydrological models, which continues to date without implementing numerical error control recognized as essential for scientific computing in other branches of science and engineering, is a major embarrassment for the hydrological community. Indeed, the findings in this study suggest that many published conclusions on parameter sensitivity, calibrated values and associated uncertainty, and, more disconcertingly, the interpretation of hydrologic models to gain insights into internal catchment dynamics, including the relative significance and behavior of different processes, may be questionable due to numerical artifacts introduced by unreliable time stepping schemes. The computational savings bought by omitting numerical quality control are dubious and cannot be justified, especially given that a multitude of robust and efficient time stepping schemes are mature, quite easy to implement, and widely available through numerical tool kits. We hope that the vivid empirical findings of this study, backed by decades of solid and uncontroversial applied mathematics, will motivate the hydrological community to address its numerical problems.

Appendix A: Sobol-Saltelli Sensitivity Analysis

[127] The global parameter sensitivity is computed using the *Saltelli* [2002] implementation of the Sobol' method [*Sobol'*, 1993].

[128] Consider the following vectors of model performance indices for N_{TGS} parameter sets,

$$\boldsymbol{\Psi} = [\Psi(\boldsymbol{\theta}_1), \Psi(\boldsymbol{\theta}_2), \ldots \Psi(\boldsymbol{\theta}_{N_{\mathrm{TGS}}})] \tag{A1}$$

$$\boldsymbol{\Psi}^{-j} = \left[\Psi\left(\boldsymbol{\theta}_1^{-j}\right), \Psi\left(\boldsymbol{\theta}_2^{-j}\right), \ldots \Psi\left(\boldsymbol{\theta}_{N_{\mathrm{TGS}}}^{-j}\right)\right], \tag{A2}$$

where $\Psi(\boldsymbol{\theta})$ and $\Psi(\boldsymbol{\theta}^{-j})$ denote the model performance indices for unperturbed and perturbed parameter sets, $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{-j}$, respectively. Note that the subscripts on $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$, and $\boldsymbol{\theta}_{N_{\mathrm{TGS}}}$ in (A1)–(A2) are used to index the parameter sets rather than individual parameters within a given set.

[129] In equations (A1)–(A2), the perturbed and unperturbed values are defined, for the rth parameter set, as

$$\boldsymbol{\theta}_r = \left[\theta_{r,1}, \theta_{r,2}, \ldots \theta_{r,k}\right] \tag{A3}$$

$$\boldsymbol{\theta}_r^{-j} = \left[\theta_{r,1}, \theta_{r,2}, \ldots, \theta_{r,(j-1)}, \theta'_{r,j}, \theta_{r,(j+1)}, \ldots \theta_{r,k}\right], \tag{A4}$$

where the prime indicates which parameter is perturbed, e.g., $\theta'_{r,j}$ indicates that the jth parameter of the rth set is perturbed.

[130] The total sensitivity of the model to its jth parameter, S_j^{TOT}, is then defined as

$$\widehat{S}_j^{\mathrm{TOT}} = 1 - \frac{\widehat{U}_{-j} - \widehat{E}^2(\Psi)}{\widehat{V}(\Psi)}, \tag{A5}$$

where $\widehat{E}^2$ and $\widehat{U}_{-j}$ are, respectively, a "squared mean" and a "perturbed" variance of a selected index of model behavior (here, the RMSE (7)), computed from N_{TGS} perturbed and unperturbed parameter sets, and $\widehat{V}$ is the total variance. These quantities are defined below:

$$\widehat{U}_{-j}(\Psi) = \frac{1}{N_{\mathrm{TGS}} - 1} \sum_{r=1}^{N_{\mathrm{TGS}}} \Psi_r \times \Psi_r^{-j} \tag{A6}$$

$$\widehat{E}^2(\Psi) = \frac{1}{N_{\mathrm{TGS}}} \sum_{r=1}^{N_{\mathrm{TGS}}} \Psi_r^2, \tag{A7}$$

$$\widehat{V}(\Psi) = \frac{1}{N_{\mathrm{TGS}}} \sum_{r=1}^{N_{\mathrm{TGS}}} \Psi_r^2 - \left[\frac{1}{N_{\mathrm{TGS}}} \sum_{r=1}^{N_{\mathrm{TGS}}} \Psi_r\right], \tag{A8}$$

where the subscript r denotes an individual parameter set.

[131] In this work the total global sensitivity indices S_j^{TOT} were calculated using $N_{\mathrm{TGS}} = 10{,}000$ parameter sets, sampled using the quasi-random Sobol' sequence [*Bratley and Fox*, 1988]. The same parameter samples were used to calculate S_j^{TOT} for all time stepping schemes, ensuring that any differences in S_j^{TOT} are caused solely by differences in the numerical model implementation.

[132] **Acknowledgments.** We are grateful to Ross Woods for the MARVEX data and to Yun Duan for the MOPEX data. We also thank Daniel Collins, Richard Ibbitt, Bethanna Jackson, George Kuczera, Hilary McMillan, and Mark Thyer for useful discussions, and the three reviewers, including Jasper Vrugt, for their insightful comments. This work was funded by the New Zealand Foundation for Research Science and Technology (contract CO1X0812), the National Aeronautic and Space Administration (contract NNG06GH10G), and the National Oceanic and Atmospheric Administration (contract NA06OAR4310065).

References

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 Resour. Res.*, *37*(4), 937–947, doi:10.1029/2000WR900363.

Bates, D. M., and D. G. Watts (1988), *Nonlinear Regression Analysis and Its Applications*, John Wiley, New York.

Behrangi, A., B. Khakbaz, J. A. Vrugt, Q. Y. Duan, and S. Sorooshian (2008), Comment on "Dynamically dimensioned search algorithm for computationally efficient watershed model calibration" by Bryan A. Tolson and Christine A. Shoemaker, *Water Resour. Res.*, *44*, W12603, doi:10.1029/2007WR006429.

Beven, K. (2008), *Environmental Modelling: An Uncertain Future*?, Taylor and Francis, London.

Beven, K., and A. Binley (1992), The future of distributed models—Model calibration and uncertainty prediciton, *Hydrol. Processes*, *6*(3), 279–298, doi:10.1002/hyp.3360060305.

Box, G. E. P., and G. C. Tiao (1992), *Bayesian Inference in Statistical Analysis*, John Wiley, New York.

Bratley, P., and B. L. Fox (1988), Implementing Sobol's quasirandom sequence generator, *Trans. Math. Software*, *14*(1), 88–100, doi:10.1145/42288.214372.

Clark, M. P., and D. Kavetski (2010), The ancient numerical demons of conceptual hydrological modeling: 1. Fidelity and efficiency of time stepping schemes, *Water Resour. Res.*, *46*, W10511, doi:10.1029/2009WR008894.

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 Resour. Res.*, *44*, W00B02, doi:10.1029/2007WR006735.

Coleman, T., M. A. Branch, and A. Grace (2006), *Optimization Toolbox for Use With MATLAB: User's Guide*, Mathworks, Natick, Mass.

Conn, A. R., N. I. M. Gould, and P. L. Toint (2000), *Trust Region Methods*, SIAM, Philadelphia, Pa.

Demidenko, E. (2000), Is this the least squares estimate?, *Biometrika*, *87*(2), 437–452, doi:10.1093/biomet/87.2.437.

Dennis, J. E., Jr., and R. B. Schnabel (1996), *Numerical Methods for Unconstrained Optimization and Nonlinear Equations, SIAM, Philadelphia, Pa.*

Duan, Q. Y., S. Sorooshian, and V. Gupta (1992), Effective and efficient global optimization for conceptual rainfall-runoff models, *Water Resour. Res.*, *28*(4), 1015–1031, doi:10.1029/91WR02985.

Duan, Q. Y., S. Sorooshian, and V. K. Gupta (1994), Optimal use of the SCE-UA global optimization method for calibrating watershed models, *J. Hydrol.*, *158*(3–4), 265–284, doi:10.1016/0022-1694(94)90057-4.

Duan, Q., et al. (2006), Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops, *J. Hydrol.*, *320*(1–2), 3–17, doi:10.1016/j.jhydrol.2005.07.031.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2003), *Bayesian Data Analysis*, 2nd ed., CRC Press, London.

Gill, P. E., W. Murray, and M. Wright (1981), *Practical Optimization*, Academic, London.

Gupta, H. V., T. Wagener, and Y. Q. Liu (2008), Reconciling theory with observations: Elements of a diagnostic approach to model evaluation, *Hydrol. Processes*, *22*(18), 3802–3813, doi:10.1002/hyp.6989.

Gupta, V. K., and S. Sorooshian (1985), The automatic calibration of conceptual catchment models using derivative-based optimization algorithms, *Water Resour. Res.*, *21*(4), 473–485, doi:10.1029/WR021i004p00473.

Hendrickson, J. D., S. Sorooshian, and L. E. Brazil (1988), Comparison of Newton-type and direct search algorithms for calibration of conceptual rainfall-runoff models, *Water Resour. Res.*, *24*, 691–700, doi:10.1029/WR024i005p00691.

Ibbitt, R., and R. Woods (2004), Rescaling the topographic index to improve the representation of physical processes in catchment models, *J. Hydrol.*, *293*(1–4), 205–218, doi:10.1016/j.jhydrol.2004.01.016.

Ivanov, V. Y., E. R. Vivoni, R. L. Bras, and D. Entekhabi (2004), Catchment hydrologic response with a fully distributed triangulated irregular network model, *Water Resour. Res.*, *40*, W11102, doi:10.1029/2004WR003218.

Kahaner, D., C. Moler, and S. Nash (1989), *Numerical Methods and Software*, Prentice Hall, Englewood Cliffs, N. J.

Kavetski, D., and G. Kuczera (2007), Model smoothing strategies to remove microscale discontinuities and spurious secondary optima in objective functions in hydrological calibration, *Water Resour. Res.*, *43*, W03411, doi:10.1029/2006WR005195.

Kavetski, D., S. Franks, and G. Kuczera (2003a), Confronting input uncertainty in environmental modelling, in *Calibration of Watershed Models*, *Water Sci. Appl.*, vol. 6, edited by Q. Duan et al., pp. 49–68, AGU, Washington, D. C.

Kavetski, D., G. Kuczera, and S. W. Franks (2003b), Semidistributed hydrological modeling: A "saturation path" perspective on TOPMODEL and VIC, *Water Resour. Res.*, *39*(9), 1246, doi:10.1029/2003WR002122.

Kavetski, D., G. Kuczera, and S. W. Franks (2006a), Calibration of conceptual hydrological models revisited: 1. Overcoming numerical artefacts, *J. Hydrol.*, *320*(1–2), 173–186, doi:10.1016/j.jhydrol.2005.07.012.

Kavetski, D., G. Kuczera, and S. W. Franks (2006b), Calibration of conceptual hydrological models revisited: 2. Improving optimisation and analysis, *J. Hydrol.*, *320*(1–2), 187–201, doi:10.1016/j.jhydrol.2005.07.013.

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

Kavetski, D., G. Kuczera, M. Thyer, and B. Renard (2007), Multistart Newton-type optimization methods for the calibration of conceptual hydrological models, in *MODSIM 2007 International Congress on Modelling and Simulation*, edited by L. Oxley and D. Kulasiri, pp. 2513–2519, Modell. and Simul. Soc. of Aust. and N. Z., Canberra.

Kirchner, J. W. (2006), Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology, *Water Resour. Res.*, *42*, W03S04, doi:10.1029/2005WR004362.

Klemes, V. (1986), Operational testing of hydrological simulation models, *Hydrol. Sci. Bull.*, *31*, 13–24, doi:10.1080/02626668609491024.

Konikow, L. F., and J. D. Bredehoeft (1992), Groundwater models cannot be validated, *Adv. Water Resour.*, *15*(1), 75–83, doi:10.1016/0309-1708(92)90033-X.

Koren, V., M. Smith, and Q. Duan (2003), Use of a Priori parameter estimates in the derivation of spatially consistent parameter sets of rainfall-runoff models, in *Calibration of Watershed Models*, *Water Sci. Appl.*, vol. 6, pp. 239–254, edited by Q. Duan et al., AGU, Washington, D. C.

Kuczera, G., and S. W. Franks (2002), Testing hydrologic models: Fortification or falsification?, in *Mathematical Modelling of Large Watershed Hydrology*, edited by V. P. Singh and D. K. Frevert, pp. 141–16, Water Resour. Publ., Littleton, Colo.

Kuczera, G., and E. Parent (1998), Monte Carlo assessment of parameter uncertainty in conceptual catchment models: The Metropolis algorithm, *J. Hydrol.*, *211*, 69–85, doi:10.1016/S0022-1694(98)00198-X.

Lambert, J. D. (1991), *Numerical Methods for Ordinary Differential Systems: The Initial Value Problem*, John Wiley, New York.

Madsen, H., G. Wilson, and H. C. Ammentrop (2002), Comparison of different automated strategies for calibration of rainfall-runoff models, *J. Hydrol.*, *261*(1–4), 48–59, doi:10.1016/S0022-1694(01)00619-9.

Merz, R., and G. Blöschl (2004), Regionalisation of catchment model parameters, *J. Hydrol.*, *287*(1–4), 95–123, doi:10.1016/j.jhydrol.2003.09.028.

Miller, D. A., and R. A. White (1999), A conterminous United States multilayer soil characteristics data set for regional climate and hydrology modeling, *Earth Interact.*, *2*, 1–26.

Nelder, J. A., and R. Mead (1965), A simplex method for function minimization, *Comput. J.*, *7*, 308–313.

Nocedal, J., and S. J. Wright (1999), *Numerical Optimization*, doi:10.1007/b98874, Springer, New York.

Pappenberger, F., K. J. Beven, M. Ratto, and P. Matgen (2008), Multimethod global sensitivity analysis of flood inundation models, *Adv. Water Resour.*, *31*(1), 1–14, doi:10.1016/j.advwatres.2007.04.009.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (1992), *Numerical Recipes in Fortran 77: The Art of Scientific Computing*, 2nd ed., Cambridge Univ. Press, Cambridge, U. K.

Reichert, P., M. Schervish, and M. J. Small (2002), An efficient sampling technique for Bayesian inference with computationally demanding models, *Technometrics*, *44*(4), 318–327, doi:10.1198/004017002188618518.

Renard, B., D. Kavetski, G. Kuczera, M. Thyer, and S. W. Franks (2010), Understanding predictive uncertainty in hydrologic modeling: Le challenge of identifying input and structural errors, *Water Resour. Res.*, *46*, W05521, doi:10.1029/2009WR008328.

Saltelli, A. (2002), Making best use of model evaluations to compute sensitivity indices, *Comput. Phys. Commun.*, *145*(2), 280–297, doi:10.1016/S0010-4655(02)00280-1.

Schoups, G., J. A. Vrugt, F. Fenicia, and N. C. van de Giesen (2010), Inaccurate numerical implementation of conceptual hydrologic models corrupts accuracy and efficiency of MCMC simulation, *Water Resour. Res.*, doi:10.1029/2009WR008648, in press.

Shampine, L. F. (1994), *Numerical Solution of Ordinary Differential Equations*, Chapman and Hall, New York.

Shampine, L. F., and M. W. Reichelt (1997), The MATLAB ODE suite, *SIAM J. Sci. Comput.*, *18*(1), 1–22, doi:10.1137/S1064827594276424.

Shampine, L. F., S. Thompson, J. A. Kierzenka, and G. D. Byrne (2005), Non-negative solutions of ODEs, *Appl. Math. Comput.*, *170*(1), 556–569, doi:10.1016/j.amc.2004.12.011.

Shoemaker, C. A., R. G. Regis, and R. C. Fleming (2007), Watershed calibration using multistart local optimization and evolutionary optimization with radial basis function approximation, *Hydrol. Sci. J.*, *52*(3), 450–465, doi:10.1623/hysj.52.3.450.

Skahill, B. E., and J. Doherty (2006), Efficient accommodation of local minima in watershed model calibration, *J. Hydrol.*, *329*(1–2), 122–139, doi:10.1016/j.jhydrol.2006.02.005.

Sobol', I. M. (1993), Sensitivity analysis for nonlinear mathematical models, *Math. Mod. Comput. Exp.*, *1*, 407–414.

Tait, A., and R. Woods (2007), Spatial interpolation of daily potential evapotranspiration for New Zealand using a spline model, *J. Hydrometeorol.*, *8*(3), 430–438, doi:10.1175/JHM572.1.

Thyer, M., G. Kuczera, and B. C. Bates (1999), Probabilistic optimization for conceptual rainfall-runoff models: A comparison of the shuffled complex evolution and simulated annealing algorithms, *Water Resour. Res.*, *35*(3), 767–773, doi:10.1029/1998WR900058.

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

Tolson, B. A., and C. A. Shoemaker (2007), Dynamically dimensioned search algorithm for computationally efficient watershed model calibration, *Water Resour. Res.*, *43*, W01413, doi:10.1029/2005WR004723.

Tolson, B. A., and C. A. Shoemaker (2008), Reply to comment on "Dynamically dimensioned search algorithm for computationally efficient watershed model calibration" by Ali Behrangi et al., *Water Resour. Res.*, *44*, W12604, doi:10.1029/2008WR006862.

van Werkhoven, K., T. Wagener, P. Reed, and Y. Tang (2008), Characterization of watershed model behavior across a hydroclimatic gradient, *Water Resour. Res.*, *44*, W01429, doi:10.1029/2007WR006271.

Vrugt, J. A., and B. A. Robinson (2007), Improved evolutionary optimization from genetically adaptive multimethod search, *Proc. Natl. Acad. Sci. U. S. A.*, *104*(3), 708–711, doi:10.1073/pnas.0610471104.

Vrugt, J. A., H. V. Gupta, W. Bouten, and S. Sorooshian (2003), A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters, *Water Resour. Res.*, *39*(8), 1201, doi:10.1029/2002WR001642.

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

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

Wagener, T., K. van Werkhoven, P. Reed, and Y. Tang (2009), Multiobjective sensitivity analysis to understand the information content in streamflow observations for distributed watershed modeling, *Water Resour. Res.*, *45*, W02501, doi:10.1029/2008WR007347.

Wolpert, D. H., and W. G. Macready (1997), No free lunch theorems for optimization, *IEEE Trans. Evol. Comput.*, *1*, 67–82, doi:10.1109/4235.585893.

Wood, W. L. (1990), *Practical Time-Stepping Schemes*, Oxford Univ. Press, New York.

Woods, R. A., R. B. Grayson, A. W. Western, M. J. Duncan, D. J. Wilson, R. I. Young, R. P. Ibbitt, R. D. Henderson, and T. A. McMahon (2001), Experimental design and initial results from the mahurangi river variability experiment: MARVEX, in *Observations and Modelling of Land Surface Hydrological Processes*, edited by V. Lakshmi et al., pp. 201–213, AGU, Washington, D. C.

Yatheendradas, S., T. Wagener, H. Gupta, C. Unkrich, D. Goodrich, M. Schaffner, and A. Stewart (2008), Understanding uncertainty in distributed flash flood forecasting for semiarid regions, *Water Resour. Res.*, *44*, W05S19, doi:10.1029/2007WR005940.

M. P. Clark, National Center for Atmospheric Research, PO Box 3000, Boulder, CO 80307-3000, USA. (mclark@ucar.edu)

D. Kavetski, Environmental Engineering, University of Newcastle, Callaghan, NSW 2308, Australia.