Groundwater

Practical Use of Computationally Frugal Model Analysis Methods

by Mary C. Hill[1,2], Dmitri Kavetski[3], Martyn Clark[4], Ming Ye[5], Mazdak Arabi[6], Dan Lu[7], Laura Foglia[8], and Steffen Mehl[9]

Abstract

Three challenges compromise the utility of mathematical models of groundwater and other environmental systems: (1) a dizzying array of model analysis methods and metrics make it difficult to compare evaluations of model adequacy, sensitivity, and uncertainty; (2) the high computational demands of many popular model analysis methods (requiring 1000's, 10,000 s, or more model runs) make them difficult to apply to complex models; and (3) many models are plagued by unrealistic nonlinearities arising from the numerical model formulation and implementation. This study proposes a strategy to address these challenges through a careful combination of model analysis and implementation methods. In this strategy, computationally frugal model analysis methods (often requiring a few dozen parallelizable model runs) play a major role, and computationally demanding methods are used for problems where (relatively) inexpensive diagnostics suggest the frugal methods are unreliable. We also argue in favor of detecting and, where possible, eliminating unrealistic model nonlinearities—this increases the realism of the model itself and facilitates the application of frugal methods. Literature examples are used to demonstrate the use of frugal methods and associated diagnostics. We suggest that the strategy proposed in this paper would allow the environmental sciences community to achieve greater transparency and falsifiability of environmental models, and obtain greater scientific insight from ongoing and future modeling efforts.

Introduction

Mathematical models are critical to addressing many environmental problems. Models embody the many types of data and knowledge typical of environmental systems, and represent known constraints related to conservation of mass and energy, population dynamics, and so on. Models are used to understand consequences of, for example, management decisions, weather variability, and climate change on environmental hazards, sustainability, and resource allocation. The growing importance and visibility of models is concurrent with increased scrutiny and criticism as the political process copes with increasing strain and degradation of environmental systems by human activity (e.g., Brodbeck 2012). As part of this development, issues of uncertainty and risk are becoming more prominent in the scientific and social discourse (e.g., Beven 2009; Smith and Stern 2011; Tartakovsky 2013; Herman et al. 2014), the public discussion is becoming more heated (e.g., Dicks 2013), and the related political decisions are becoming more difficult and controversial (e.g., Interacademy Council [IAC] 2010).

Responding to such intense scrutiny and criticism of models requires greater clarity about how models represent real processes and relate to data–that is, it requires greater transparency and falsifiability (Popper 1959; Carrera and Neuman 1986, Oreskes et al. 1994; Oreskes and Belitz 2001; Beven 2009; Clark et al. 2011; Gupta et al. 2012). Models of environmental systems are rarely, if ever, unique (e.g., Beven 2009), yet even nonunique models can provide insight about environmental systems. By revealing the sensitivity of model predictions to data and process representation, as required for transparency, the realism of simulated relations can be evaluated. By testing models against data, as required for falsifiability, models gain legitimate credibility and trust.

Three Challenges

Over the last decades, as models of environmental systems and high-performance computing have evolved,

[1]Corresponding author: Department of Geology, University of Kansas, 1475 Jayhawk Blvd., Lawrence, KS 66049; 785-864-2728; mchill@ku.edu
[2]U.S. Geological Survey, Boulder, CO
[3]University of Adelaide, Adelaide, Australia; dmitri.kavetski@adelaide.edu.au
[4]National Center for Atmospheric Research, Boulder, CO; mclark@ucar.edu
[5]Florida State University, Tallahassee, FL; mye@fsu.edu
[6]Colorado State University, Fort Collins, CO; marabi@engr.colostate.edu
[7]Oak Ridge National Laboratory, Oak Ridge, TN; lud1@ornl.gov
[8]University of Darmstadt, Darmstadt, Germany; foglia@geo.tu-darmstadt.de
[9]California State University, Chico, CA; smehl@csuchico.edu
Received April 2014, accepted February 2015.

doi: 10.1111/gwat.12330

we suggest that three challenges have arisen that compromise the transparency and falsifiability of environmental models:

(1) *A "Tower of Babel" of model analysis methods*. The large and seemingly ever-growing number of analysis methods proposed for the development and evaluation of mathematical models (e.g., see Figure 1) causes confusion and impedes a clear dialog between scientists, modelers, and decision makers (also noted by Pappenberger and Beven 2006). Even when different methods measure the same aspect of model analysis, results are often presented in ways that make implications difficult to compare. These factors impede the ability to conduct transparent model comparisons and evaluations.

(2) *"Colossal" computational burdens*. Many modern models and model analysis methods require Goliath-size computational resources. For example, Sobol' sampling and Markov-chain Monte Carlo (MCMC) model analysis methods typically require 1000s, 10,000s, or more model runs to provide a thorough exploration of the model parameter space (e.g., Razavi et al. 2010; Herman et al. 2013a). Despite increasing computer capabilities and the expansion of parallel computing, near exclusive reliance on computationally demanding model analysis methods is impractical in fields where forward model runs can sometimes take days or more to complete. Furthermore, environmental models themselves are increasingly computationally expensive as modelers strive for higher spatial and temporal resolution, larger domain size, more realistic process representation, more defined parameters, and by considering more alternative models to investigate model adequacy (e.g., Hunt et al. 2007; Doherty and Welter 2010; Clark et al. 2011; Wood et al. 2011; Foglia et al. 2013; Herman et al. 2013b; Hill et al. 2013). Although in principle the computational burden can be reduced using surrogate models and emulators, which are often constructed statistically (e.g., Razavi et al. 2012), the emulators themselves may be expensive to construct, the construction needs to be at least partially repeated for each alternative model, and it is generally difficult to ascertain whether the emulators can reproduce the relevant nuances of model behavior.

Faced with computationally overwhelming model analysis tasks, modelers are often forced to: (1) simplify—and possibly oversimplify!—models merely to reduce execution times, and/or (2) conduct analyses with fewer model runs than needed to obtain reliable results. Both choices are detrimental to using models to test alternative hypotheses and to quantify uncertainty, leading to reduced falsifiability and transparency. Model simplification merely to enable a particular type of analysis also undermines the efforts invested by model developers to identify and represent more accurately important processes and characteristics of environmental systems. Finally, the computational demand impedes the replication of results, which is essential for transparency and falsifiability.

(3) "Numerical daemons." Many environmental models contain "artificial nonlinearities" that reduce model realism and complicate model analysis. For example, thresholds are often used to simulate processes that in reality are likely to be smooth at the spatio-temporal scales relevant to the model (Kavetski and Clark 2010; references cited therein, and Appendix S1 [Supporting Information]). Figure 2 shows that thresholds can produce very erratic response surfaces, which complicates some types of model analysis. Smoothing of the threshold relation, as shown in the inset of Figure 2B, is often more realistic and yields smooth response surfaces that make the model much easier to evaluate. Thresholds are common in environmental models. Indeed, even the popular MODFLOW groundwater flow model includes the kind of thresholds that produce the erratic surface shown in Figure 2A, and the equation for the MODFLOW Drain Package is essentially identical. This challenge is a difficulty with models rather than with model analysis methods. We include it here because it makes model analysis difficult and because it can confound the analysis of real system nonlinearities.

In this study, we contend that some aspects of contemporary model development, analysis, and application need to be reconsidered in order to address systematically the aforementioned key challenges. The following section proposes a general strategy to achieve this.

The Strategy

The three challenges listed in the previous section can be addressed using a strategy that combines a range of model analysis methods and perspectives, and includes diagnostic tests to ensure that the methods are used appropriately.

The challenge of the model analysis "Tower of Babel" is approached using a three-pronged line of attack. First, we introduce a typology to organize the diversity of model analysis methods. The typology is shown in Figure 1, and presents a novel organization of a wide range of methods based on typical questions relevant to the analysis of model adequacy, sensitivity, uncertainty, and risk. Second, we encourage theoretical and empirical investigation of how the methods relate to each other, thus extending the efforts of Oakley and O'Hagan (2004), Borgonovo (2006), Pappenberger et al. (2006), Tang et al. (2007), Foglia et al. (2007), Kleijnen (2010), Lu et al. (2012), Li et al. (2013), and so on. Third, we suggest the need for comparable metrics and presentation of results even when different analysis methods are used, to avoid confusion when investigating questions such as those posed in Figure 1.

The challenge of "Colossal" computational demands for model analysis can be tackled by providing a choice of methods that take few to many model runs. Methods that take few model runs (10s to 100s) are called computationally frugal; methods that take many model runs (1000s, 10,000s, and more) are called computationally demanding (Figure 1). Frugal methods include Newton-type optimization methods and first-order sensitivity analysis and uncertainty intervals. Some frugal methods work well only if solutions are smooth with respect to

Common questions	Frugal methods	Demanding methods
Model Adequacy		
1. How can many data types with variable quality be included?	Error-based weighting and SOO or MAP	MOO, Pareto curve
2. Is model misfit/overfit a problem? Is the fit to prior knowledge and data subsets consistent? Are errors Gaussian?	RMSE, Nash-Sutcliffe, graphs, R^2_N, s_n^2, $s_{(n-p)}^2$ Compare fit to a priori error analysis using s_n^2, $s_{(n-p)}^2$	MOO, Pareto curve
3. How nonlinear is the problem?	Intrinsic nonlinearity, DELSA	DELSA, Explore objective function
Sensitivity and Uncertainty		
Observations (Obs) ↔ Parameters (Pars)		
4. What pars can and cannot be estimated with the obs?	Scaled local stats (CSS, ID, PCC, etc.), SVD, DoE, MoM(OAT, EE)	DoE, MoM(OAT, EE), eFAST, Sobol', RSA
5. Are any parts dominated by one obs and, thus, its error?	Scaled local stats (Leverage, DFBETAS)	Cross validation
6. How certain are the par values?	Par uncertainty intervals	Par uncertainty intervals
7. Which obs are important and unimportant to pars?	Scaled local stats (Leverage, Cook's D)	Cross validation
Parameters (Pars) ↔ Prediction (Preds)		
8. Which pars are important and unimportant to preds?	Scaled local stats (PSS, etc.), DELSA	DELSA, eFAST, Sobol'
9. How certain are the preds?	z/SD_z, Pred uncertainty intervals	Pred uncertainty intervals, multi-model analysis
10. Which pars contribute most and least to the pred uncertainty?	Scaled local stats (PPR VOII)	eFAST, Sobol'
Observations (Obs) ↔ Prediction (Preds)		
11. Which existing and potential obs are important to preds?	Scaled local stats (OPR VOII)	Cross validation
12. For multi-model analysis, which models are likely to produce accurate preds?	Analyze model fit and estimated parameters, AIC, AICc, BIC, KIC	Cross validation
Risk Assessment		
13. What risk is associated with a given decision strategy and set of scenarios?	Combine uncertainty analysis and scenario simulation. Smooth cost function	Combine uncertainty analysis and scenario simulation. Cost function need not be smooth.
14. What are decisions are robust given a set of uncertain scenarios?	Evolutionary multiobjective optimization. Within this demanding method use frugal model analysis methods.	

Figure 1. Questions of interest when modeling environmental systems. For each question selected, computationally frugal and demanding model analysis methods are listed. Some methods can be conducted such that they are either frugal or demanding, and are listed in both columns. Sensitivity and uncertainty questions are organized by whether they address the relation between "observations" and parameters, parameters and predictions, or "observations" and predictions. Here, "observations" refers to simulated values to which the observations are compared in goodness of fit measures.

parameters and some require the more restrictive assumptions of linearity and near-Gaussian errors. Applicability of computationally frugal methods needs to be tested, and the next section of this article discusses such tests. Conversely, computationally demanding methods are typically based on Monte Carlo replication or other forms of sampling simulated results over a defined parameter space; restrictive assumptions are limited to applicability of chosen parameter value ranges and conducting sufficient sampling. Lu et al. (2012) discuss tradeoffs between computational requirements and limiting assumptions of model analysis methods.

Finally, the challenge of "numerical daemons" can be removed or reduced through the use of robust model implementations, including numerically stable time stepping techniques, smoothed constitutive relationships and boundary conditions, and so on (van Genuchten 1980; Kavetski and Clark 2010; Sheets et al. 2014). In addition to improving model stability and accuracy, numerically robust models can be designed to avoid unnecessary nonsmoothness in the models response surface, facilitating the use of computationally frugal methods (Kavetski and Kuczera 2007).

Computationally frugal model analysis methods are included in or affected by all three challenges. They are some of the methods involved in the "Tower of Babel," as shown in Figure 1. They address the challenge of "Collossal" computational demands by providing less demanding alternatives. A primary limitation of many computationally frugal methods—the requirement of solution smoothness—is addressed by any reduction of "Numerical Daemons." Thus, the practical use of computationally frugal methods addressed in this article is intertwined with the strategy presented.

Frugal methods are of interest because of the following. Most fundamentally, recent work suggests that when compared to computationally demanding methods, frugal methods can provide similar insights (and sometimes additional insights) with a fraction (2% or less) of the model runs (Samarov 1993; Anderman et al. 1996; Barth and

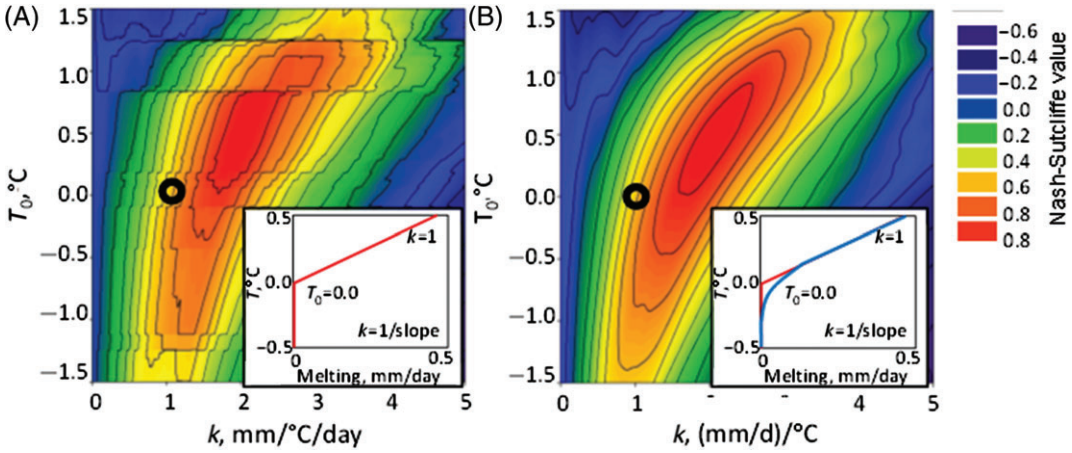


Figure 2. Effects of numerical errors on the calibration of hydrological models, investigated using Nash-Sutcliffe response surfaces for snowmelt models where (A) melting begins sharply at 0 °C vs. (B) the same model but using a smoothed melting curve likely to be more realistic for typical simulations. The circle in the contour plots identifies the values of parameters T_0 and k represented in the insets (after Kavetski and Kuczera 2007).

Hill 2005; Foglia et al. 2007; Kavetski and Clark 2010; Rakovec et al. 2014). As a result, frugal methods make it easier to analyze complex models (Foglia et al. 2009; Hill et al. 2013), to compare systematically multiple alternative models (Foglia et al. 2013; Rakovec et al. 2014), and to gain insight into the models used for computationally demanding evaluation of management options (e.g., Herman et al. 2014).

Important remaining issues include diagnostic tests for identifying when results from frugal methods are reliable, suggestions for selecting between computationally frugal and demanding methods, and illustrations of using computationally frugal methods. These are addressed in the remainder of the study.

Diagnostic Tests for Computationally Frugal Methods

Many frugal methods are based on assumptions of model smoothness (or the stronger assumption of linearity), Gaussian error behavior, and(or) a single minimum within the range of plausible parameter values and model fit (Hill and Tiedeman 2007; Saltelli et al. 2008). Some frugal methods are referred to as "local" because they describe model behavior in narrow regions of the parameter space. This includes all measures with "local" in the name used in Figure 1 and the definitions provided in the "Notation" section.

Local methods are strongly affected by solution roughness. They rely on model behavior (including model derivatives) evaluated at a single parameter set. When derivatives are used, they are either calculated analytically or, much more frequently, approximated using parameter perturbation methods, and their values can change dramatically for small parameter values changes when the response surface is rough, as in Figure 2A.

To establish whether frugal method results are meaningful, tests are conducted to determine whether the model is too nonlinear or rough, and the error too non-Gaussian for the frugal methods to retain useful accuracy.

The diagnostic methods are presented in the context of uncertainty intervals calculated using three methods (Lu et al. 2012). The intervals are shown in Figure 3A and include intervals produced using computationally frugal to demanding methods. Diagnostics for model linearity, Gaussian errors, and adequacy are reported in Figure 3B; they are calculated at optimized parameter values with observation weighting (see Appendix S1). The results presented suggest the following.

1 *Analysis of model linearity*. The intrinsic nonlinearitity measure determines a set of parameters defined at the edges of the 95% parameter confidence region based on linear theory, and compares linearized and full model results calculated at those values. Results in Figure 3B suggest significant, but not extreme nonlinearity with respect to the observed quantities.
2 *Analysis of Gaussian error*. R_N^2 indicates that overall the weighted residuals are Gaussian, which suggests that the total errors are Gaussian. R_N^2 is the correlation between standard normal deviates and ordered weighted residuals. The Kolmogorov–Smirnov statistic could be used instead, but has a higher chance of erroneously identifying non-Gaussian distributions as Gaussian.
3 *Analysis of model adequacy*. The $s_{(n-p)}$ value close to 1.0 for model INT correctly identifies it as the only model with negligible error not accounted for by the error-based weighting (refer to Appendix S1 for more on error-based weighting). For this problem, the error-based weighting was assigned using only an analysis of data error, so the value close to 1.0 indicates that the model error is negligible relative to the data error. Plots of weighted residuals and simulated values (not shown; see Lu et al. 2012) show consistent results: they suggest that models HO and 3Z are biased and favor model INT as being more accurate.

The analysis in Figure 3 was conducted at the optimized parameter values. The utility of frugal methods also needs to be evaluated when model development begins, and for problems lacking optimal values. For these circumstances, derivative-based measures calculated for different parameter values can be used. For example, if the parameters identified as important and unimportant by local statistic composite-scaled sensitivity (CSS) change radically as the model is run with sets of reasonable parameter values that are close together, a rough (nonsmooth) response or objective-function surface is indicated and frugal methods will likely not be useful for the problem as posed. In addition to using routine model runs conducted in any model calibration effort, derivative-based tests for roughness can also be considered more formally using approaches such as those found in the PEST-related program SENSAN (Doherty 2010) or the new DELSA method (Rakovec et al. 2014). DELSA is now available in the Sensitivity R Package (Joseph Guillaume, 2015, Aalto University, Finland, written commun.)

Selecting Between Frugal and Demanding Model Analysis Methods

Selecting between frugal and expensive methods requires consideration of the strengths and weaknesses of both types of methods. The previous section discussed difficulties with frugal methods and relevant diagnostics. The results of computationally demanding methods can be inaccurate and/or meaningless if an insufficient number of model runs is carried out to achieve convergence, parameter limits are unfortunately defined, or the average values global methods produce are misleading (Rakovec et al. 2014). Furthermore, global optimization methods may succeed in finding the global optimum, but more realistic parameter values that produce slightly inferior optima are rarely reported so that users may overlook information highly relevant to model adequacy and realism (Kavetski and Clark 2010). As such, no methods can be assumed a priori to work well for a given problem; assessment is possible through careful posterior diagnostics.

A model analysis procedure consistent with the strategy presented in this work is illustrated in Figure 4. First, diagnostic tests are applied to check for problems such as solution roughness. As discussed in the last section of this article, some of these tests involve computationally frugal sensitivity analysis methods. If difficulties are detected, the results of the preliminary tests can often guide the user in the application of more computationally demanding methods. Using this procedure, the cost of the initial frugal analysis is a mere fraction of the total analysis cost, and is frequently accompanied by valuable insights. If, however, the diagnostic tests support the assumptions of the frugal methods, the computationally demanding methods need not be invoked at all, thus saving the modeler a tremendous computational effort.

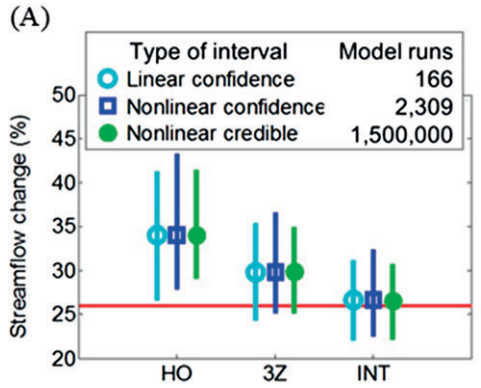
Examples of Judiciously Applying Computationally Frugal Model Analysis Methods

The effectiveness of our strategy depends critically on the ability to select and combine computationally frugal and demanding methods, and on diagnostic tests to investigate whether a particular analysis method is being used appropriately. This section illustrates these procedures using a selection of case studies from groundwater and surface water hydrology.

The first example was used previously to demonstrate use of diagnostic tests. Here it is used to illustrate an uncertainty analysis that involves using computationally frugal to demanding methods (linear and nonlinear confidence intervals and nonlinear credible intervals) (Lu et al. 2012). Figure 3A shows uncertainty intervals on the prediction of flow to a stream given increased groundwater pumpage calculated with three nonlinear groundwater models (models HO, 3Z, and INT). For each model, the results are similar in terms of the interval inaccuracy caused by model bias (interval distance from the true values, which are known for this synthetic problem) and in terms of identifying more precise models (model INT has the smallest intervals). For this problem, the differences between the three alternative models are large compared to the differences in the three types of confidence intervals. The implication that substantial model nonlinearity may have small consequences on frugal methods relative to the differences produced by alternative models may be applicable to a wide range of environmental modeling problems. In such circumstances, using frugal model analysis methods allows greater exploration of alternative models and hence greater transparency and falsifiability than may be possible using computationally demanding model analysis methods.

The second example illustrates sensitivity analysis of a simple six-parameter synthetic groundwater flow test case. Figure 5 compares sensitivity analysis results from computationally frugal and demanding methods and illustrates how metrics can be presented in a way that facilitates comparison, even when different methods of analysis are used. Figure 5B shows a plot of first-order effects that is readily compared to the computationally frugal CSS results shown in Figure 5C; Figure 5B and C identifies the same most and least important parameters, and reporting the sum in Figure 5B maintains the information about parameter interactions. The ability to achieve comparability is an argument for using graphs such as that shown in Figure 5B to display first-order effects.

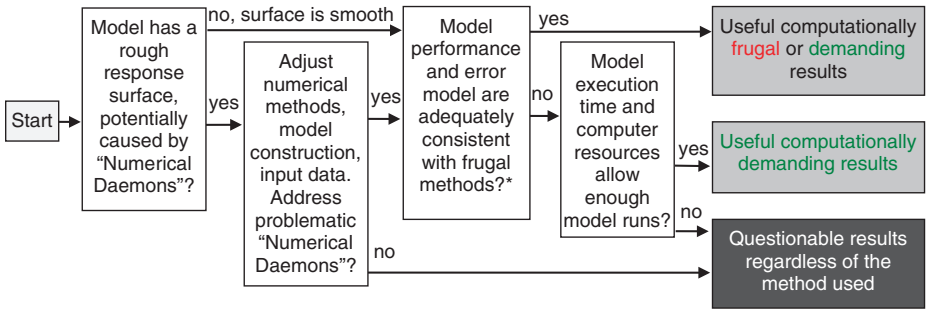
Figure 5D and E shows how parameter interactions can be quantified for individual parameters using computationally demanding and frugal methods, respectively. Figure 5D shows an unusual presentation of what are commonly computationally demanding first- and total-order effects. Figure 5E shows an unusual presentation of frugal parameter correlation coefficient (PCC) statistics. For Figure 5D and E, the data are generally presented as a



(B)

Model	HO	3Z	INT
Test model adequacy using $s^2_{(n-p)}$			
	1.49	1.27	1.05
	(1.25-1.84)	(1.06-1.57)	(0.88-1.30)
Test model linearity using intrinsic nonlinearity			
	0.54	0.04	0.18
Test for Gaussian independent weighted residuals using			
	0.989(0.96)	0.986(0.96)	0.989(0.97)

Figure 3. (A) Results of computationally frugal and demanding uncertainty quantification methods. (B) Results of computationally frugal tests for model adequacy, model nonlinearity, and Gaussian errors. Darker shading identifies are less ideal results and greater likely advantage of the computationally demanding methods. Results are reported for three models: HO has a homogeneous hydraulic-conductivity distribution, 3Z uses three zones of constant value, and INT uses interpolation. Confidence intervals (in parentheses) on the bias-corrected error variance, $s^2_{(n-p)}$, indicate that none of the models exhibit over-fitting (entire interval less than 1.0) and HO and 3Z exhibit model inadequacy not accounted for by the error-based weighting (entire interval larger than 1.0). Intrinsic nonlinearity indicates that HO and INT are nonlinear (values between 0.09 and 1.0) and 3Z is moderately nonlinear (0.01–0.09); greater nonlinearity theoretically means poorer performance of the linear intervals. R^2_N (critical values are in parentheses) indicates that all residuals are Gaussian and independent (modified from Lu et al. 2012).



* Yes: Simulated results vary smoothly as parameter values change for parameter values of interest. If multiple local minima exist one can be identified as the realistic solution. The error model is unimodal. These characteristics are determined based onknowledge of model theory and numerical methods, and (or)easily conducted diagnostic tests.

Figure 4. Flowchart showing how three major issues affect the utility of computationally frugal and demanding analysis methods: characteristics of the process-model, the error model, and available computer resources.

table and matrix of numbers, respectively; this graphical presentation clearly identifies the interrelated and correlated parameters. Figures 5D and E identify the same parameters as interacting too strongly to permit unique identification. Sampling methods eFAST (10,182 model runs; results shown in Figure 5), Sobol' (15,400), and the Method of Morris (70) all identify the same parameter characteristics. The local method CSS required 13 fully parallelizable model runs. The results were evaluated in the context of a related two-parameter objective function (not shown), which was smooth. In this case, the differences shown in Figure 5 reflect the distributed values provided by the local methods and the point values produced by global methods. The similarity in contrast to the erratic and misleading local results that would be produced in the objective function was rough (e.g., see Figure S1A in Appendix S1, and other examples presented by Kavetski and Clark 2010).

The third example illustrates the application of frugal methods to computationally expensive models. In this case study, the rainfall-runoff model TOPKAPI (Liu et al. 2005) was used to simulate the 160 km^2 Maggia basin in southern Switzerland for which one 4-month forward run typically took 40 min. The available computer resources were adequate for hundreds and perhaps a few thousand model runs. Error-based weighting and single objective function (SOO) were used (see Figure 1; the SOO objective function was the sum of squared weighted residuals, with error-based weights defined as described by Foglia et al. 2009). Figure 6A shows computationally frugal measures of individual parameter sensitivity (here, CSS) and parameter correlation (here, PCC). This information was used to identify four parameters as important and also interdependent; ultimately three of these were estimated using prior information. In the calibrated model, a bias-corrected variance, $s^2_{(n-p)}$, value of 15.9 suggests that

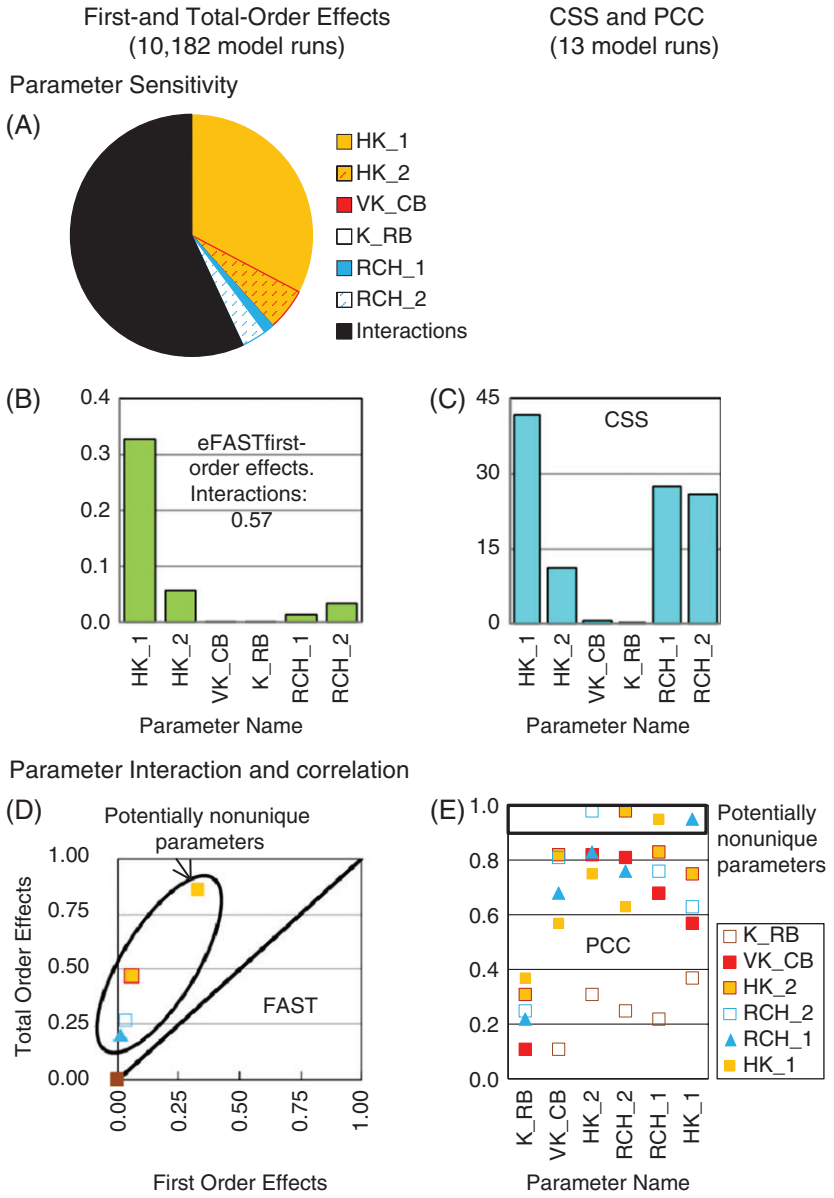


Figure 5. Parameter sensitivity, interaction, and correlation for the simple groundwater problem of Hill and Tiedeman (2007). (A, B, D) Global method first- and total-order effects (calculated with eFAST using the sum of squared weighted residuals objective function as the performance measure). (C, E) Local method CSS and PCC statistics. (A) A common way to present first-order effects and a quantification of parameter interactions. (B, C) Easily comparable presentations for individual parameter sensitivity. Larger values indicate more important individual parameters. Colors are coordinated with Figure 1 (green for a computationally demanding global method and red for a frugal method). Parameters that interact or are correlated enough that it is likely to prevent unique estimation of parameters are identified by (D) total-order effects larger than first-order effects (dots above the line) and (E) absolute values of PCC close to 1.00. We believe that (D) and (E) illustrate new ways to analyze and illustrate parameter interactions and correlations.

model misfit exceeds what would be consistent with the weighting by about a factor of 4 ($15.9^{1/2}$). This suggests the presence of model error beyond that accounted for in the weight calculation. Systematic differences between observed and simulated flows in Figure 6B provide additional information about the model bias. The Cook's *D* metric (Cook and Weisberg 1982), is a measure of observation importance derived given the assumption of a linear model. Cook's *D* was calculated at the estimated parameter values (Figure 6B) and indicates the sensitivity of the estimated parameter values to each observation. In this problem, Cook's *D* revealed unexpectedly important low-flow observations, invalidating an a priori assumption that the information provided by the large flood peak overwhelms the information provided by low flows. It then follows that the low flows will require denser sampling prior to further modeling efforts. Given the computational cost of the rainfall-runoff model, this insight would be infeasible to obtain using expensive analysis methods, but was readily obtained using the frugal methods based on just 196 model runs for the entire analysis.

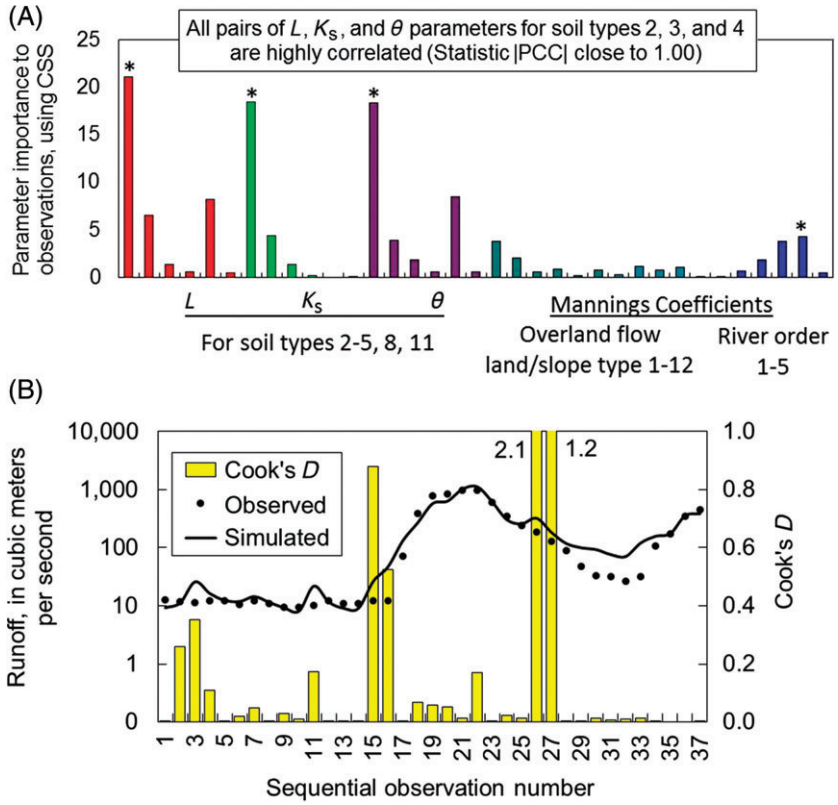


Figure 6. Results from computationally frugal model sensitivity analysis of a TOPKAPI rainfall-runoff model of the Maggia basin in southern Switzerland. (A) Important parameters and their interactions as measured by CSS and PCC. The 35 parameters include depth (L), saturated hydraulic conductivity (K_s), and water content (θ) for five soil types, and Manning's coefficients for overland flow and rivers. Asterisk (*) identifies optimized parameters; the interactions of soil type 2 parameters required prior information for reasonable estimation to be achieved. Results shown are evaluated at the starting parameter values; similar results are obtained at intermediate and final parameter values. (B) Model fit to observations following optimization, and observation importance to the set of defined parameters estimated using a frugal method. The analysis required 71 highly parallelizable model runs at starting and estimated parameter values (142 model runs total). Best-fit parameters were obtained using 54 model runs, of which four sets of eight could be parallelized (modified from Foglia et al. 2009).

Conclusions

The ability of mathematical models to provide insights into environmental systems and predict their future behavior depends critically on the transparency of the modeling process and on the ability to test rigorously ("falsify") the models. However, the growing multitude of model analysis methods—which we liken to a "Tower of Babel"—means that studies are being conducted using a plethora of methods and metrics related in ways that are poorly understood. Moreover, increasing computer capabilities notwithstanding, we argue that near-exclusive reliance on computationally "colossal" methods is unsustainable in a field where forward model runs can take anywhere from seconds to months to complete. Finally, many existing models are implemented using approaches that introduce significant spurious artifacts and unrealistic nonlinearities—"numerical daemons"—making many models more nonlinear than the real world. These spurious effects greatly, and unnecessarily, complicate model analysis.

Our suggested strategy confronts these three challenges as follows. The "Tower of Babel" challenge is addressed by organizing the many available model evaluation methods and metrics based on function and computational demand. By including computationally frugal to demanding methods and encouraging exploration of functional similarities despite theoretical and algorithmic differences, the strategy promotes methods appropriate to the wide range of applications typical of environmental problems, and thus tempers the demand for "colossal" computational efforts. In particular, diagnostic tests consisting largely of a limited computationally frugal analysis can indicate whether or not computationally demanding methods are needed. The cost of the early diagnostic is then a mere fraction of the total cost, and is frequently accompanied by valuable insights. Finally, we suggest that identifying and eradicating "numerical daemons" in environmental models is a critical problem that poses a grand challenge to future model development.

Three case studies are used to illustrate the opportunities available through the proposed strategy, including identification of important parameters, parameter interactions and correlations, observation dominance in complex models, and uncertainty of estimated parameters and

simulated predictions. It is our hope that the increase in falsifiability and transparency achievable through the approach outlined in this work will contribute to a more productive, exciting, and societally consequential future for environmental modeling.

Acknowledgments

Mary Hill and Steffen Mehl were funded by the U.S. Geological Survey programs National Water Quality Assessment (NAWQA), Groundwater Resources Program (GWRP), and National Research Program (NRP). Ming Ye and Dan Lu were funded by NSF-EAR grant 0911074 and the DOE Early Career Award DE-SC0008272. Laura Foglia was funded by Swiss National Science Foundation (SNF) grant number 21–66885. We thank George Kuczera (University of Newcastle) and Sujay Kumar (NASA) for comments on early versions of the manuscript. We are also grateful to the editors of Ground Water and three anonymous reviewers.

Notations

AIC, AICc, BIC, and KIC	Model discrimination criteria (Poeter and Hill 2007; Foglia et al. 2013; and references cited therein)
Cook's *D*	Identifies the actual importance of the parameters to observations. For this linear measure, the importance depends on the observed value. Potential importance is measured by leverage (Cook and Weisberg 1982)
Cross-validation	Methods for which observation importance is measured by removing one or more observations and repeating the analysis (Good 2001; Foglia et al. 2007, 2013)
CSS	Composite scaled sensitivity, a scaled local statistic
DELSA	Distributed Evaluation of Local Sensitivity Analysis (Rakovec et al. 2014). A local measure is calculated for many sets of parameter values. Can be applied to any local measure
DFBETAS	Identifies the actual importance of each observation to each parameter
DoE	Design of Experiment using fractional factorial methods (Montgomery 2012)
EE	Elementary effects (Saltelli et al. 2008)
Error-based weighting	Weighting based on an analysis of errors prior to any simulations, and sometimes updated based on model simulations. Error-based weights can be used to include many data types in a single objective function (SOO), producing a commonly computationally frugal method of estimating parameters relative to MOO. See Appendix S1 for additional information Described by Reed et al. (2012)
Evolutionary multi-objective optimization	
Explore objective function	Create plots such as in Figure 2. UCODE_2014 (Poeter et al. 2014) can produce the needed data sets
FAST, eFAST	Fourier amplitude sensitivity testing (Cukier et al. 1978; Saltelli et al. 1999)
Intrinsic nonlinearity	Measure of model nonlinearity (Bates and Watts 1980; Cooley 2004)
Leverage	Identifies observations that could potentially be important to parameters. For this linear measure, the actual importance depends on the observed values and is measured by Cook's D (Cook and Weisberg 1982; Helsel and Hirsch 2002; Hill and Tiedeman 2007)
MAP	Maximum a posteriori (Oliver et al. 2008)
MoM	Method of Morris (Saltelli et al. 2008)
MOO	Multi-objective optimization (Deb 2001). See the Appendix S1 for a discussion of MOO and SOO
OAT	One at a time parameter sampling (Saltelli et al. 2008)
OPR	Observation–prediction statistic for value of information (Tiedeman et al. 2004; Tonkin et al. 2007). Dausman et al. (2010) discuss a similar statistic
PPR	Parameter-prediction statistic for value of information (Tiedeman et al. 2003; Tonkin et al. 2007)
RMSE	Root mean-squared error
RSA	Regionalized sensitivity analysis (Spear and Hornberger 1980)
Scaled local statistics	(CSS, ID, PCC, Leverage, Cook's *D*, DFBETAS, PSS, etc.) (Cook and Weisberg 1982;

	Helsel and Hirsch 2002; Hill and Tiedeman 2007; Doherty and Hunt 2009; 2010; Hill 2010)
SCE	Shuffled complex evolution (Duan et al. 1992)
SOO	Single-objective optimization (Hill and Tiedeman 2007). See the Appendix S1 for a discussion of SOO and MOO
SVD	Singular value decomposition used to reparameterize the model (Tonkin and Doherty 2005; Aster et al. 2013; Poeter et al. 2014)
Uncertainty intervals	Can be calculated using linear, nonlinear, MCMC, null-space Monte Carlo, and bootstrapping methods. Linear methods are computationally frugal. Some of the others can be very expensive (Chernick 2007; Lu et al. 2012; and references cited therein)
z/SD_z	A t-statistic on the prediction z (Draper and Smith 1998, 128, 31, 125–127; Hill and Tiedeman 2007)
σ_n^2	Sum of squared, possibly weighted, residuals (observed minus simulated values) divided by the number of members in the sum (n) minus the number of estimated parameters ($n-p$). Division by ($n-p$) produces a bias-corrected result (Draper and Smith 1998) When weights are used, this is a bias-corrected, weight-standardized variance.
$\sigma_{(n-p)}^2$	Sum of squared, possibly weighted, residuals divided by the number of members in the sum. Residuals are observed minus simulated values

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1. Appendix containing four problems of model analysis and an outline of the strengths, weaknesses, and compatibilities of frugal methods in tackling these problems.

References

Anderman, E.R., M.C. Hill, and E.P. Poeter. 1996. Two-dimensional advective transport in ground-water flow parameter estimation. *Ground Water* 34, no. 6: 1001–1009.

Aster, R.C., B. Borschers, and C.H. Thurber. 2013. *Parameter Estimation and Inverse Problems*. Waltham, Massachusetts: Academic Press.

Barth, G.R., and M.C. Hill. 2005. Parameter and observation importance in modeling virus transport in saturated systems–Investigations in a homogenous system. *Journal of Contaminant Hydrology* 80: 107–129.

Bates, D.M., and D.G. Watts. 1980. Relative curvature measures of nonlinearity. *Journal of the Royal Statistical Society, Series B* 42, no. 1: 1–25.

Beven, K. 2009. *Environmental Modeling, An Uncertain Future*. Abingdon, UK: Rutledge.

Borgonovo, E. 2006. Measuring uncertainty importance, investigation and comparison of alternative approaches. *Risk Analysis* 26: 1349–1361.

Brodbeck, N. 2012. World Risk Report 2012. Berlin, Germany: Alliance Development Works. http://www.ehs.unu.edu/file/get/10487.pdf (accessed March 2, 2015).

Carrera, J., and S.P. Neuman. 1986. Estimation of aquifer parameters under transient and steady state-conditions. *Water Resources Research* 22, no. 2: 199–343.

Chernick, M.R. 2007. *Bootstrapping Methods*. New York, New York: Wiley.

Clark, M.P., D. Kavetski, and F. Fenicia. 2011. Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resources Research* 47: W09301. DOI:10.1029/2010WR009827.

Cook, R.D., and S. Weisberg. 1982. *Residuals and Influence in Regression*. New York: Chapman & Hall.

Cooley, R.L. 2004. *A theory for modeling ground-water flow in heterogeneous media. U.S. Geological Survey Professional Paper 1679*. Reston, Virginia: USGS.

Cukier, R.I., H.B. Levine, and K.E. Schuler. 1978. Nonlinear sensitivity analysis of multiparameter model systems. *Journal of Computational Physics* 26: 1–42.

Dausman, A.M., J. Doherty, C.D. Langevin, and M.C. Sukop. 2010. Quantifying data worth toward reducing predictive uncertainty. *Ground Water* 48, no. 5: 729–740.

Deb, K. 2001. *Multiobjective Optimization using Evolutionary Algorithms*. New York: Wiley.

Dicks, L. 2013. Bees, lies and evidence-based policy. *Nature* 494, no. 7437: 283.

Doherty, J., and R.J. Hunt. 2009. Two statistics for evaluating parameter identifiability and error reduction. *Journal of Hydrology* 366: 119–127.

Doherty, J., and R.J. Hunt. 2010. Response to comment on "Two statistics for evaluating parameter identifiability and error reduction." *Journal of Hydrology* 380: 489–496.

Doherty, J., and D. Welter. 2010. Short exploration of structural noise. *Water Resources Research* 46: W05525. DOI:10.1029/2009WR008377.

Draper, N.R., and H. Smith. 1998. *Applied Regression Analysis*, 3rd ed. Hoboken, New Jersey: Wiley.

Duan, Q., S. Sorooshian, and V. Gupta. 1992. Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resources Research* 28, no. 4: 1015–1031.

Foglia, L., S.W. Mehl, M.C. Hill, P. Perona, and P. Burlando. 2007. Testing alternative ground water models using cross validation and other methods. *Ground Water* 45, no. 5: 627–641. DOI:10.1111/j.1745-6584.2007.00341.x.

Foglia, L., M.C. Hill, S.W. Mehl, and P. Burlando. 2009. Sensitivity analysis, calibration, and testing of a distributed hydrological model using error-based weighting and one objective function. *Water Resouces Research* 45: W06427. DOI:10.1029/2008WR007255.

Foglia, L., S.W. Mehl, M.C. Hill, and P. Burlando. 2013. Evaluating model structure adequacy: The case of the Maggia Valley groundwater system, southern Switzerland. *Water Resources Research* 49. DOI:10.1029/2011WR011779.

Good, P.I. 2001. *Resampling Methods—A Practical Guide to Data Analysis*. Boston, Massachusetts: Birkhauser.

Gupta, H.V., M.P. Clark, J.A. Vrugt, G. Abramowitz, and M. Ye. 2012. Towards a comprehensive assessment of model structural adequacy. *Water Resources Research* 48. DOI:10.1029/2011WR011044.

Helsel, D.R., and R.M. Hirsch. 2002. *Statistical methods in water resources. U.S. Geological Survey Techniques of Water-Resource Investigation: 04-A3*. Reston, Virginia: USGS.

Herman, J.D., H.B. Zeff, P.M. Reed, and G.W. Characklis. 2014. Beyond optimality: Multistakeholder robustness tradeoffs for regional water portfolio planning under deep uncertainty. *Water Resources Research* 50. DOI:10.1002/2014WR015338.

Herman, J.D., J.B. Kollat, P.M. Reed, and T. Wagener. 2013a. Technical note: Method of Morris effectively reduces the computational demands of global sensitivity analysis for distributed watershed models. *Hydrology and Earth System Sciences Discussions* 10: 4275–4299. DOI:10.5194/hessd-10-4275-2013.

Herman, J.D., P.M. Reed, and T. Wagener. 2013b. Time-varying sensitivity analysis clarifies the effects of watershed model formulation on model behavior. *Water Resources Research* 49. DOI:10.1002/wrcr.20124.

Hill, M.C. 2010. Comment on "Two statistics for evaluating parameter identifiability and error reduction" by John Doherty and Randall J. Hunt. *Journal of Hydrology* 380: 481–488.

Hill, M.C., and C.R. Tiedeman. 2007. *Effective Groundwater Model Calibrations, with Analysis of Data, Sensitivities, Predictions, and Uncertainty*. New York: Wiley.

Hill, M.C., C.C. Faunt, W.R. Belcher, D.S. Sweetkind, C.R. Tiedeman, and D. Kavetski. 2013. Knowledge, transparency, and refutability in groundwater models, an example from the Death Valley regional groundwater flow system. *Journal of the Physics and Chemistry of the Earth* 64: 105–116.

Hunt, R.J., J. Doherty, and M.J. Tonkin. 2007. Are models too simple? Arguments for increased parameterization. *Ground Water* 45, no. 3: 254–261. DOI:10.1111/j.1745-6584.2007.00316.x.

Interacademy Council (IAC). 2010. *Climate Change Assessments, Review of the Processes and Procedures of the IPCC*. Intergovernmental Panel on Climate Change. http://www.ipcc.ch/pdf/IAC_report/IAC%20Report.pdf (accessed February 3, 2011).

Kavetski, D., and M.P. Clark. 2010. Ancient numerical daemons of conceptual hydrological modeling: 2, impact of time stepping schemes on model analysis and prediction. *Water Resources Research* 46: W10511. DOI:10.1029/2009WR008896.

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

Kleijnen, J.P.C. 2010. Sensitivity analysis of simulation models: An overview. 6th International conference on sensitivity analysis of model output (SAMO) 2010, Milan, Italy. *Procedia—Social and Behavioral Sciences* 2, no. 6: 7585–7586.

Li, J.D., Q.Y. Duan, W. Gong, A.Z. Ye, Y.J. Dai, C.Y. Miao, Z.H. Di, C. Tong, and Y.W. Sun. 2013. Assessing parameter importance of the common land model based on qualitative and quantitative sensitivity analysis. *Hydrology and Earth System Sciences* 10: 2243–2286. DOI:10.5194/hessd-10-2243-2013.

Liu, Z., M.L.V. Martina, and E. Todini. 2005. Flood forecasting using a fully distributed model: Application of the TOPKAPI model to the upper Xixian catchment. *Hydrology and Earth System Sciences* 9, no. 4: 347–364.

Lu, D., M. Ye, and M.C. Hill. 2012. Analysis of regression confidence intervals and Bayesian credible intervals for uncertainty quantification. *Water Resources Research* 48. DOI:10.1029/2011WR011289.

Montgomery, D.C. 2012. *Design of Experiments*, 8th ed. New York, New York: Wiley.

Oakley, J.E., and A. O'Hagan. 2004. Probabilistic sensitivity analysis of complex models: A Bayesian approach. *Journal of the Royal Statistical Society: Series B* 66, no. part 3: 751–769.

Oliver, D.S., A.C. Reynolds, and N. Liu. 2008. *Inverse Theory for Petroleum Reservoir Characterization and History Matching*. Cambridge, UK: Cambridge University Press.

Oreskes, N., and K. Belitz. 2001. Philosophical issues in model assessment. In *Model Validation: Perspectives in Hydrological Science*, ed. M.G. Anderson and P.D. Bates, 23–41. Chichester, UK: Wiley.

Oreskes, N., K. Shrader-Frechette, and K. Belitz. 1994. Verification, validation, and confirmation of numerical models in the earth sciences. *Science* 263: 641–646.

Pappenberger, F., and K.J. Beven. 2006. Ignorance is bliss: Or seven reasons not to use uncertainty analysis. *Water Resources Research* 42: W05302. DOI:10.1029/2005WR004820.

Pappenberger, F., H. Harvey, K. Beven, J. Hall, and I. Meadowcroft. 2006. Decision tree for choosing an uncertainty analysis methodology: A wiki experiment. *Hydrological Processes* 20, no. 17: 3793–3798.

Poeter, E.P., and M.C. Hill. 2007. *MMA: A computer code for multi-model analysis. U.S. Geological Survey Techniques and Methods 6-E3, 113*. Reston, Virginia: USGS.

Poeter, E.P., M.C. Hill, D. Lu, and S.W. Mehl. 2014. *UCODE_2014, with New Capabilities to Define Parameters Unique to Predictions, Calculate Weights using Simulated Values, Estimate Parameters with SVD, Evaluate Uncertainty with MCMC, and More*. Integrated Groundwater Modeling Center, Colorado School of Mines. Report Number GWMI 2014-02.

Popper, K. 1959. *The Logic of Scientific Discovery*. London, UK: Hutchinson.

Rakovec, O., M.C. Hill, M.P. Clark, A.H. Weerts, A.J. Teuling, and R. Uijlenhoet. 2014. Distributed Evaluation of Local Sensitivity Analysis (DELSA), with application to hydrologic models. *Water Resources Research* 50. DOI:10.1002/2013WR014063.

Razavi, S., B.A. Tolson, L.S. Matott, N.R. Thomson, A. MacLean, and F.R. Seglenieks. 2010. Reducing the computational cost of automatic calibration through model preemption. *Water Resources Research* 46: W11523. DOI:10.1029/2009WR008957.

Razavi, S., B.A. Tolson, and D.H. Burn. 2012. Review of surrogate modeling in water resources. *Water Resources Research* 48: W07401. DOI:10.1029/2011WR011527.

Reed, P.M., D. Hadka, J.D. Herman, J.R. Kasprzyk, and J.B. Kollat. 2012. Evolutionary multiobjective optimization in water resources: The past, present, and future. *Advances in Water Resources*. DOI:10.1016/j.advwatres.2012.01.005.

Saltelli, A., S. Tarantola, and K.P.S. Chan. 1999. A quantitative model-independent method for global sensitivity analysis of model output. *Technometrics* 41, no. 1: 39–56.

Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saosana, and S. Tarantola. 2008. *Global Sensitivity Analysis: The Primer*. New York: Wiley.

Samarov, A.M. 1993. Exploring regression structure using nonparametric functional estimation. *Journal of the American Statistical Association* 88, no. 423: 836–847.

Sheets, R.A., M.C. Hill, H.M. Haitjema, A.M. Provost, and J.P. Masterson. 2014. Simulation of water-table aquifers using specified saturated thickness. *Ground Water* 53, no. 1: 151–157. DOI:10.1111/gwat.12164.

Smith, L.A., and N. Stern. 2011. Uncertainty in science and its role in climate policy. *Philosophical Transactions of the Royal Society A* 369: 1–24. DOI:10.1098/rsta.2011.0149.

Spear, R.C., and G.M. Hornberger. 1980. Eutrophication in Peel Inlet: II, Identification of critical uncertainties via generalized sensitivity analysis. *Water Research* 14, no.b 43: 49.

Tang, Y., P. Reed, T. Wagener, and K. van Werkhoven. 2007. Comparing sensitivity analysis methods to advance lumped watershed model identification and calibration. *Hydrology and Earth System Sciences* 11: 793–817. DOI:10.5194/hess-11-793-2007.

Tartakovsky, D.M. 2013. Assessment and management of risk in subsurface hydrology: A review and perspective. *Advances in Water Resources* 51: 247–260.

Tiedeman, C.R., M.C. Hill, F.A. D'Agnese, and C.C. Faunt. 2003. Methods for using groundwater model predictions to guide hydrogeologic data collection, with application to the Death Valley regional ground-water flow system. *Water Resources Research* 39, no. 1: 5–1. DOI:10.1029/2001WR001255.

Tiedeman, C.R., D.M. Ely, M.C. Hill, and G.M. O'Brien. 2004. A method for evaluating the importance of system state observations to model predictions, with application to the Death Valley regional groundwater flow system. *Water Resources Research* 40: W12411. DOI:10.1029/2001WR001255.

Tonkin, M.J., and J. Doherty. 2005. A hybrid regularized inversion methodology for highly parameterized environmental models. *Water Resources Research* 41: W10412. DOI:10.1029/2005WR003995.

Tonkin, M.J., C.R. Tiedeman, D.M. Ely, and M.C. Hill. 2007. OPR-PPR, a computer program for assessing data importance to model predictions using linear statistics. U.S. Geological Survey, Techniques and Methods Report TM-6E2, 115. Reston, Virginia: USGS. https://water.usgs.gov/software/OPR-PPR/ (accessed March 2, 2015).

Wood, E.F., J.K. Roundy, T.J. Troy, L.P.H. van Beek, M.F.P. Bierkens, E. Blyth, A. de Roo, P. Döll, M. Ek, J. Famiglietti, D. Gochis, N. van de Giesen, P. Houser, P.R. Jaffé, S. Kollet, B. Lehner, D.P. Lettenmaier, C. Peters-Lidard, M. Sivapalan, J. Sheffield, A. Wade, and P. Whitehead. 2011. Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water. *Water Resources Research* 47: W05301. DOI:10.1029/2010WR010090.

Authors' Note

The authors do not have any conflicts of interest or financial disclosures to report.