Journal of Hydrology 366 (2009) 119–127

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Two statistics for evaluating parameter identifiability and error reduction

John Doherty [a], Randall J. Hunt [b,*]

[a] Watermark Numerical Computing, 336 Cliveden Avenue, Corinda 4075, Brisbane, Australia
[b] US Geological Survey, 8505 Research Way, Middleton, WI 53562, United States

ARTICLE INFO

Article history:
Received 1 May 2008
Received in revised form 4 December 2008
Accepted 22 December 2008

This manuscript was handled by Konstantine P. Georgakakos, Editor-in-Chief, with the assistance of Enrique R. Vivoni, Associate Editor

Keywords:
Parameter estimation
Identifiability
Error reduction
Modeling

SUMMARY

Two statistics are presented that can be used to rank input parameters utilized by a model in terms of their relative identifiability based on a given or possible future calibration dataset. Identifiability is defined here as the capability of model calibration to constrain parameters used by a model. Both statistics require that the sensitivity of each model parameter be calculated for each model output for which there are actual or presumed field measurements. Singular value decomposition (SVD) of the weighted sensitivity matrix is then undertaken to quantify the relation between the parameters and observations that, in turn, allows selection of calibration solution and null spaces spanned by unit orthogonal vectors. The first statistic presented, "parameter identifiability", is quantitatively defined as the direction cosine between a parameter and its projection onto the calibration solution space. This varies between zero and one, with zero indicating complete non-identifiability and one indicating complete identifiability. The second statistic, "relative error reduction", indicates the extent to which the calibration process reduces error in estimation of a parameter from its pre-calibration level where its value must be assigned purely on the basis of prior expert knowledge. This is more sophisticated than identifiability, in that it takes greater account of the noise associated with the calibration dataset. Like identifiability, it has a maximum value of one (which can only be achieved if there is no measurement noise). Conceptually it can fall to zero; and even below zero if a calibration problem is poorly posed. An example, based on a coupled groundwater/surface-water model, is included that demonstrates the utility of the statistics.

Introduction

The fact that only a handful of the multitude of parameters that may be utilized by an environmental model are uniquely estimable on the basis of most calibration datasets has been noted in the literature (e.g., Sorooshian and Gupta, 1983; Yeh, 1986; Beck and Halfon, 1991; Beven and Binley, 1992; Beven and Freer, 2001; Vrugt et al., 2002; Doherty and Skahill, 2006; Marcé et al., 2008). The inability to uniquely identify certain parameters results from insensitivity of model outputs corresponding to historical observations of system state to these parameters, excessive correlation with other parameters, or both. Ideally, parameters that are revealed to be non-identifiable through appropriate pre-calibration analysis should be either fixed at reasonable values, or tied to one another and estimated collectively during the model calibration process.

The intention of this paper is to introduce two easily-computed statistics through which parameters can be readily compared in terms of their ability to be uniquely estimated on the basis of an existing, or posited, calibration dataset. These statistics are based on the same concepts as those employed by mathematical regularization as a device for solution of the inverse problem of model calibration; see for example Hunt et al. (2007).

Regularization has been used as a matter of course in many industries (e.g., geophysical data analysis and image processing) for many years, and is expected to assume increasing importance in calibration of hydrologic models as they become more complex and are required to simulate the details of more sophisticated environmental processes (such as the interaction between groundwaters and surface waters). Increasing model complexity is expected to exacerbate issues of parameter identifiability, as the number of parameters that can be estimated through the model calibration process is often frustratingly low. For example, calibration of a surface-water model against a single streamflow time series is expected to identify only three to five parameters (Bevin, 1989; Jakeman and Hornberger, 1993). The term "parameter" is used here in a broad sense to denote any value assigned to a model that is uncertain, whether this is a physical property over all or part of the model domain, or even a boundary condition to which an arbitrary value is assigned, and then fixed, while other values are adjusted through the calibration process. Processing the single time series into a multi-component objective function, with each component formulated from a different system response mode (see, for example, Boyle et al., 2000; Doherty and Johnston, 2003;

* Corresponding author. Tel.: +1 608 828 9901; fax: +1 608 821 3817.
E-mail addresses: johndoherty@ozemail.com.au (J. Doherty), rjhunt@usgs.gov (R.J. Hunt).

0022-1694/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhydrol.2008.12.018

Wagener et al., 2003) may enable constraint of up to eight parameters. However this is likely to be the maximum number of parameters that can be estimated on the basis of a single streamflow time series. Nevertheless, most watershed models in everyday use are equipped with many more parameters than this. For example the much-used Sacramento Soil Moisture Accounting model (Burnash et al., 1973) employs 16 parameters; the popular Hydrologic Simulation Program Fortran (HSPF) model (Bicknell et al., 2001) and the joint groundwater/surface-water model GSFLOW (Markstrom et al., 2008) use a great deal more. Where multiple land uses, and/or subwatersheds feed the same subbasin (or even multiple subbasins), the number of possible parameters represented in a model can grow rapidly (into the thousands or more), and is usually well beyond that for which unique estimation of individual parameters is possible on the basis of historical streamflow records.

The concept of parameter identifiability is salient to both model design and usage. If the overall level of parameter identifiability is low in a particular modeling context, the worth of simulating that system with a highly complex model may be questionable (e.g., Bevin, 1989). Although such a model may be capable of simulating complex, interconnected physical and chemical processes, predictions made with this model may be highly uncertain because of limited constraints imposed on parameter values pertaining to these processes by the requirement that model outputs match historical measurements of system state (Moore and Doherty, 2005). Therefore, a simpler model with fewer parameters may serve management purposes just as well.

Some assistance is available for pre-calibration evaluation of unidentifiable (or poorly identifiable) parameters through the "composite scaled sensitivity" statistic of Hill and Tiedeman (2007). However, this statistic does not address the phenomenon of parameter correlation, whereby changes in one parameter can be offset by changes in other parameters, with the result that they can be varied in certain ratios with virtually no effect on any model output for which a corresponding field measurement exists. Furthermore, no account is taken of the level of suspected measurement and structural noise within a calibration dataset, which has a pronounced effect on the information that can actually be extracted from it through the model calibration process. Therefore, it can be difficult to make the link between composite scaled sensitivities and the need to include or exclude any particular parameter from the calibration process (e.g., Hunt et al., 2006).

More sophisticated analyses are available such as the multiobjective generalized sensitivity analysis (MOGSA) algorithm discussed by Bastidas et al. (1999), and the dynamic identifiability analysis (DYNIA) method of Wagener et al. (2003). Both methods are attractive because they are based on parameter values that attempt to span the entirety of parameter space. Hence the outcomes of these analyses are rich in information, and are relatively immune from the effects of model non-linearity on parameter estimates. However their computational burden can be very high, especially with larger numbers of parameters. Furthermore, though some information is available on correlation-incurred parameter non-uniqueness, they are not designed to specifically explore this phenomenon. As parameter numbers increase to hundreds or thousands, correlation is expected to become the dominant source of parameter non-identifiability in most models. Thus, the work here emphasizes statistics that can be computed with modest computational burden and can readily accommodate parameter correlation.

If calibration is undertaken using the Gauss Marquardt Levenberg (GML) method of parameter estimation as implemented in packages such as PEST (Doherty, 2008) and UCODE-2005 (Poeter et al., 2005), post-calibration analysis of the parameter covariance matrix, and matrices/statistics derived from it (such as correlation coefficients and eigenvectors/eigenvalues) can provide a basis for identification of problematical parameters. However, if correlation and/or insensitivity is too large, the "normal matrix" used as a basis for objective function minimization by the GML method cannot be inverted and the covariance matrix required for such analysis cannot be calculated.

In this work, two easily calculated statistics are described that can be used to characterize which model parameters are identifiable on the basis of a given calibration dataset, and which are not. These statistics can be calculated before the calibration process is actually undertaken; thus, the calibration process can be adapted to the strengths and weaknesses of the current calibration dataset so that as much information as possible is extracted from that dataset, but no attempt is made to estimate parameters for which there is no basis for estimation. Although illustrated using surface-water model parameters, these statistics can be employed in any environmental modeling context where values for at least some parameters must be estimated through calibration. The first statistic is referred to herein as "parameter identifiability" and the second is referred to as "relative parameter error reduction". These statistics have the following advantages:

1. they are easily calculated (either before, during or after the calibration process);
2. no manual parameter lumping, fixing, amalgamation or simplification is required for their computation;
3. they have an easily-understood, intuitive appeal;
4. they are based on well-established theory.

However they have the disadvantages that:

1. they rely on differentiability of model outputs with respect to adjustable parameters; and
2. they are based on linear theory.

The first of the above disadvantages may limit use of these statistics in conjunction with some models, similar to other sensitivity methods. However the second disadvantage is unlikely to invalidate their utility, because statistics such as these are intended to provide qualitative rather than quantitative insights into relative parameter estimability.

The statistics

The two statistics are defined in this section; theoretical backgrounds to these definitions are provided in the Appendix. Both statistics rely on singular value decomposition of the weighted sensitivity matrix, the elements of which express the sensitivity of each model output (for which there is a corresponding field measurement) to each parameter. Both can be calculated easily using utility software available with the PEST software suite (IDENTPAR, GENLINPRED – Doherty, 2008). As both of these statistics pertain to individual model parameters, they readily furnish a basis for parameters to be compared with each other in terms of their relative capability to be estimated independently during model calibration.

In model calibration, a useful construct is to think of parameter space as having two subspaces, the calibration solution space, and the calibration null space. The former is comprised of parameter combinations that have import to model outputs for which corresponding field measurements exist; the latter is comprised of parameter combinations that have little effect on those model outputs when superimposed on a parameter set that already calibrates the model. Because of this, the calibration dataset is depleted in information through which these latter combinations of parameters can be estimated. See Moore and Doherty (2005) for further details.

The "identifiability" f_i of parameter i can be defined as:

$$f_i = (\mathbf{V}_1\mathbf{V}_1^t)_{ii} = \mathbf{i}^t(\mathbf{V}_1\mathbf{V}_1^t)\mathbf{i} \quad (1)$$

where $\mathbf{V}_1$ is a matrix whose columns are orthogonal unit vectors that span the calibration solution space, and $\mathbf{i}$ is a unit (or basis) vector in which all elements are zero except for that pertaining to the parameter in question. (As explained in the Appendix, $\mathbf{V}_1$ can be obtained through singular value decomposition of the weighted parameter sensitivity matrix.) It can be shown that f_i is the cosine of the angle between $\mathbf{i}$ and its projection onto the calibration solution space. The value of f_i can vary between zero (indicating complete non-identifiability of the parameter in question on the basis of the current calibration dataset because the parameter lies wholly within the calibration null space), and one (indicating complete identifiability because the parameter lies wholly within the calibration solution space). Its complement, the "non-identifiability" (g_i) of a parameter is defined as:

$$g_i = 1 - f_i \quad (2)$$

Although relatively easy to compute, f_i and g_i provide an incomplete characterization of the potential error associated with estimates of individual parameters. That is, if a parameter possesses an identifiability of one, this does not mean that it can be estimated with zero error. Rather, it means that any errors associated with its estimation are a product solely of noise associated with the measurement dataset rather than with non-zero projection of the parameter onto the calibration null space. The "relative parameter error reduction" statistic seeks to rectify this inadequacy.

The "relative error" (e_i) of parameter i is defined as:

$$e_i = [\sigma_2^2]_i/[\sigma_1^2]_i \quad (3)$$

where $[\sigma_2^2]_i$ is the post-calibration error variance associated with estimation of parameter i and $[\sigma_1^2]_i$ is its pre-calibration error variance. Pre-calibration error variance is the potential for error associated with assignment of a value to a parameter purely on the basis of expert knowledge of that parameter. Post-calibration error variance can also be calculated, and is computed on the basis of a notional calibration exercise of a linear model undertaken using truncated singular value decomposition as a regularization device. Such an approach provides parameter estimates of minimum norm (and hence maximum likelihood where parameters are scaled by their pre-calibration error variance and statistically independent). See Moore and Doherty (2005) and the Appendix for more details. Like f_i, e_i has a minimum value of zero. However, its maximum value can be greater than one because it is possible for errors to be magnified rather than diminished through calibration. This is an outcome of the $\mathbf{S}_1^{-1}$ component of the second term of Eq. (A9); one obvious use of this statistic is to seek prior warning of this possibility.

As for f_i, a complementary term that we call "relative parameter error reduction" (r_i) can be defined as:

$$r_i = 1 - e_i \quad (4)$$

Like the identifiability parameter f_i, this has a maximum value of one. Ideally, its minimum value is zero, but with careless formulation of the inverse problem it can fall below zero for reasons stated above.

As is apparent from the Appendix, more information is required for calculation of e_i and r_i than for calculation of f_i and g_i. This includes an estimate of the "innate variability" of each model parameter or, in other words, the potential for error in assignment of a value to the parameter based on pre-calibration expert knowledge of it. This information does not need to be exact, for it is the purpose of statistics such as these to provide indicators of parameter status rather than exact characterization of their uncertainty.

Computation of the statistics discussed above requires that sensitivities of model outputs with respect to adjustable parameters be calculated. It does not require that the actual values of calibration data, or of model parameters themselves, be known. Thus, it is an easy matter to recompute them with presumed additional calibration data, or different sets of presumed calibration data, to judge the worth of additional data in increasing parameter identifiability and/or of decreasing potential parameter error. In this regard, the statistics presented here form a more coherent basis for calculation of the worth of future data collection than statistics such as OPR-PPR described by Tonkin et al. (2007) because they explicitly include the calibration null space component of uncertainty. This component is commonly the dominant contributor to parameter and predictive uncertainty (Moore and Doherty, 2005).

An example

The two statistics are demonstrated using a groundwater/surface-water model constructed for the Trout Lake watershed in northern Wisconsin, USA, an area characterized by many surface-water features (Fig. 1). Groundwater flows through an aquifer consisting of 40–60 m of unconsolidated Pleistocene glacial outwash sands and gravels. Runoff is primarily generated locally (the hydrologic response units (HRUs) in Fig. 1 do not encompass large areas), and surface-water flows through an immature drainage network of streams and lakes. Movement of groundwater within the watershed has been previously simulated using a number of models, including an analytic element screening model and a three-dimensional, finite-difference model. Recently, movement of both surface and ground waters has been simulated using the coupled groundwater/surface-water code GSFLOW Markstrom et al., (2008), which is an integration of the United States Geological Survey codes PRMS Leavesley et al. (1983) and MODFLOW2005 (Harbaugh, 2005). The application of GSFLOW to the Trout Lake watershed has been described by Hunt et al. (2008a). The reader is referred to Walker and Bullen (2000) for a complete description of the watershed, and to Hunt et al. (2006) and Hunt et al. (2008b) for descriptions of previous groundwater modeling efforts and calibration datasets.

In this example the number of surface-water model parameters that can be identified on the basis of a calibration dataset comprised only of daily stream flows recorded at a single gaging site was investigated. The site is situated at the pour point of the North Creek subwatershed (Fig. 1) of the Trout Lake watershed; this subwatershed was chosen because of the absence of any contribution to streamflow (and related confounding factors) by upstream lakes. Eighteen GSFLOW surface-water parameters (Table 1) applied uniformly across the entire subwatershed were investigated for identifiability. The eighteen parameters include those associated with solar radiation, precipitation, snowpack, evapotranspiration, surface runoff, and soil-zone processes simulated by the coupled GSFLOW model (Fig. 2). Focus was placed on surface-water processes for the following reasons: (a) simulation of surface-water processes is often responsible for the large increase in parameters needed to run a coupled model; and (b) the results of the investigation could be compared with those of Bevin (1989) and Jakeman and Hornberger (1993) discussed above who also explored watershed model parameter identifiability. Other parameters employed by the coupled GSFLOW model (including parameters pertinent to the groundwater model component) were set to reasonable values based on previous calibration exercises; see Pint (2002), Hunt and Doherty (2006), and Hunt et al. (2008a,b) for further details.

In the present case identifiability and relative parameter error variance reduction were computed for the log of each parameter

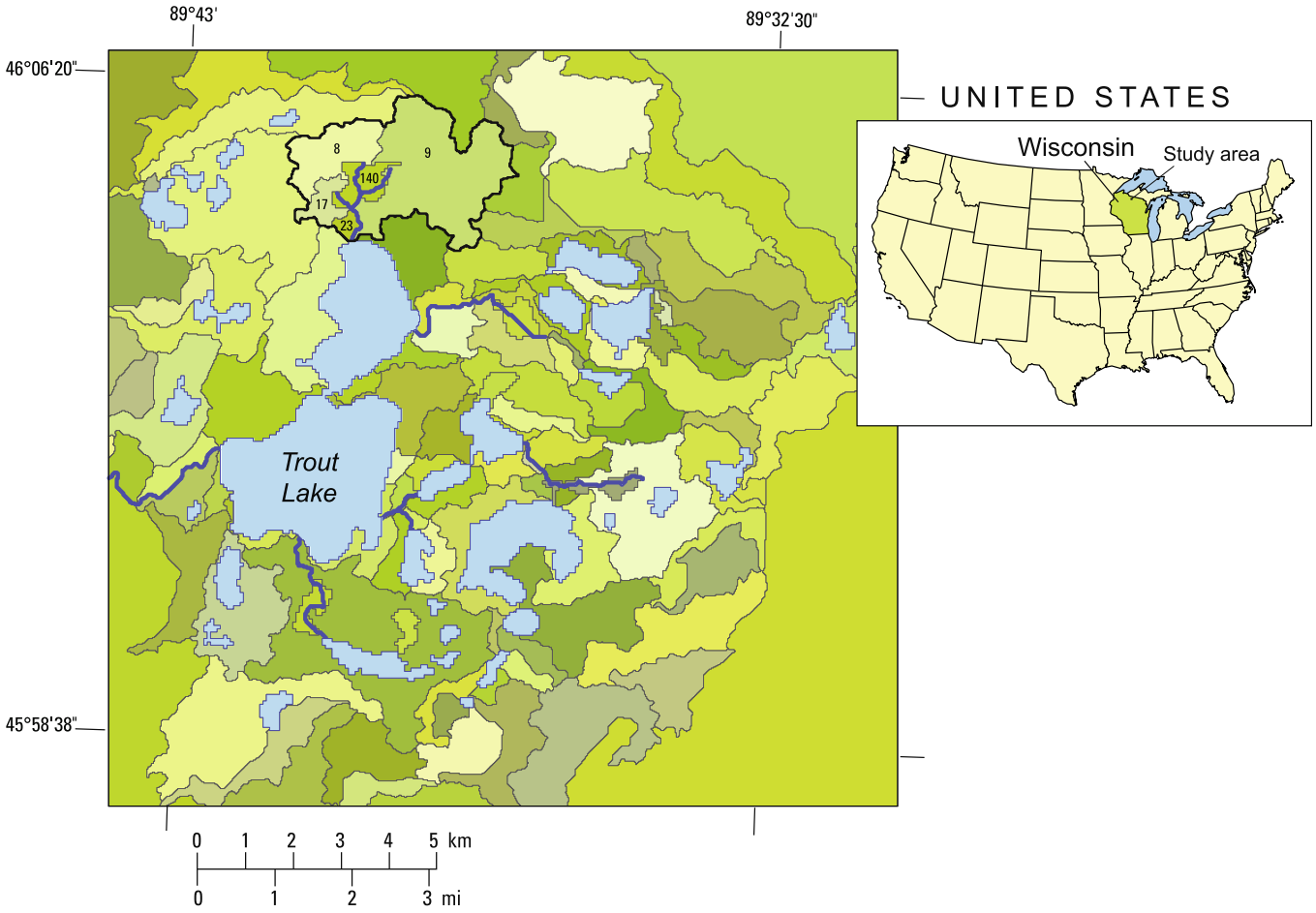

Fig. 1. Map of hydrological response units (HRUs – in green) and surface-water features (in blue) used in the Trout Lake GSFLOW model. The North Creek subwatershed used for parameter identifiability and relative parameter error reduction determinations (represented by HRUs 8, 9, 140, 17, and 23) is highlighted.

Table 1
Description of surface-water parameters used in identifiability analysis.

Name	GSFLOW module	Description	Units	Value
crad_coef	Solar radiation	Constant used in the cloud-cover to solar radiation relation	Dimensionless	0.7
crad_exp	Solar radiation	Exponent used in the cloud-cover to solar radiation relation	Dimensionless	0.5
jh_coef	Potential evapotranspiration	Monthly air temperature coefficient used in Jensen-Haise potential evapotranspiration equation	Temperature units	0.007
jh_coef_hru	Potential evapotranspiration	Air temperature coefficient used in Jensen-Haise potential evapotranspiration equation for each HRU	Temperature units	20
transp_tmax	Potential evapotranspiration	Maximum temperature used to determine when transpiration begins in an HRU	Degree-day	500
adjmix_rain	Precipitation/snow computation	Monthly adjustment factor for a mixed precipitation event as a decimal fraction	Dimensionless	1.1
den_init	Precipitation/snow computation	Density of new-fallen snow as a decimal fraction	Dimensionless	0.0983
freeh2o_cap	Precipitation/snow computation	Free-water holding capacity of snowpack expressed as decimal fraction of total snowpack water equivalent	Dimensionless	0.01
tmax_allsnow	Precipitation/snow computation	Monthly maximum air temperature at which precipitation is all snow for the HRU	Temperature units	1.90
smidx_coef	Surface runoff	Coefficient in non-linear contributing area algorithm	Dimensionless	0.01
smidx_exp	Surface runoff	Exponent in non-linear contributing area algorithm	Per inch	0.3
snowinfil_max	Surface runoff	Daily maximum snowmelt infiltration for the HRU	Inches	1
pref_flow_den	Soilzone	Decimal fraction of the soil-zone available for preferential flow	Dimensionless	0.1
slowcoef_lin	Soilzone	Linear flow-routing coefficient for slow interflow	Per day	0.015
slowcoef_sq	Soilzone	Non-linear flow-routing coefficient for slow interflow	Per inch-day	0.1
ssr2gw_exp	Soilzone	Exponent in the equation used to compute gravity drainage to PRMS ground-water reservoir or MODFLOW finite-difference cell	Dimensionless	1
ssr2gw_rate	Soilzone	Linear coefficient in the equation used to compute gravity drainage to PRMS ground-water reservoir or MODFLOW finite-difference cell	Inches per day	0.5
ssrmax_coef	Soilzone	Maximum amount of gravity drainage to PRMS ground-water reservoir or MODFLOW finite-difference cell	Inches	1

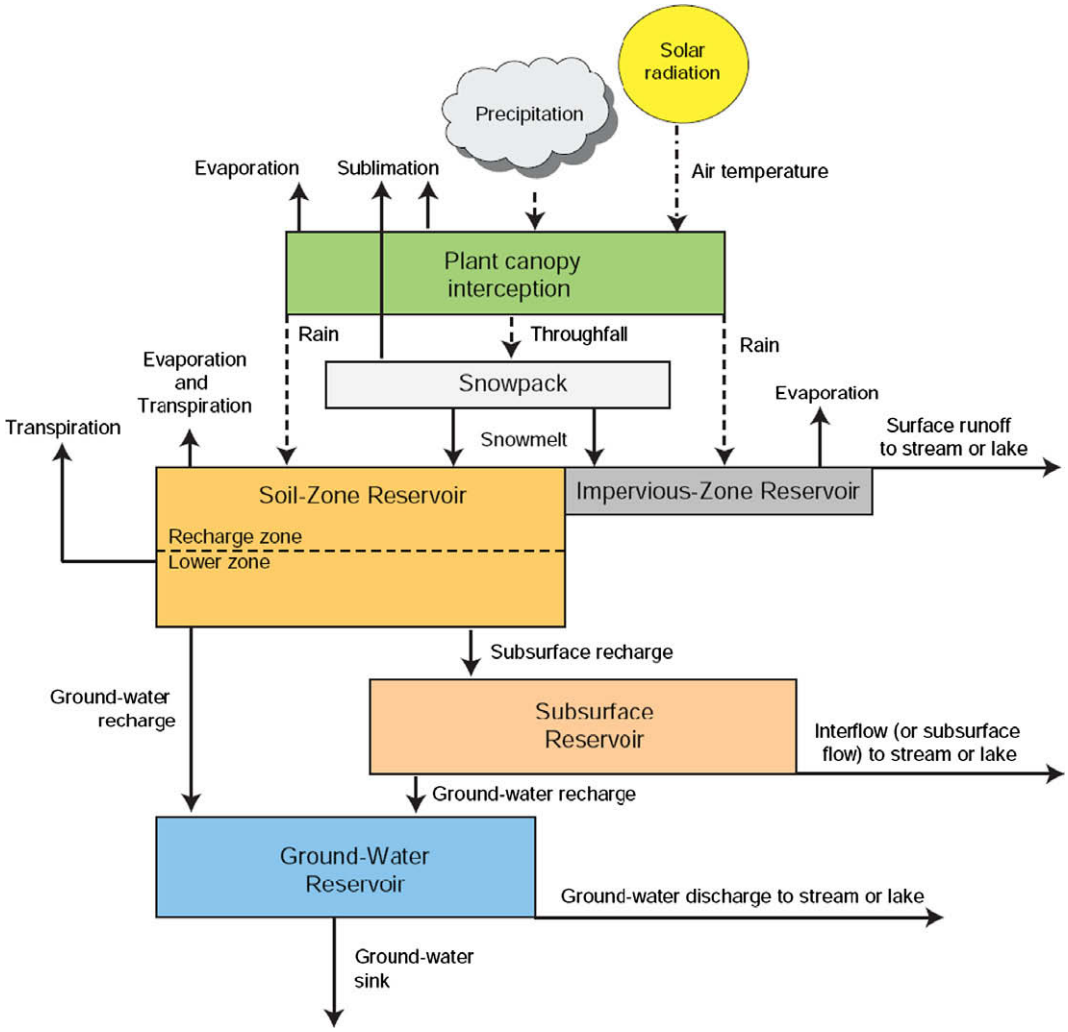

Fig. 2. Schematic of GSFLOW compartments used for surface-water modeling. Parameters involving solar radiation, precipitation, snowpack, evapotranspiration, surface runoff, and the soil-zone are investigated in this work.

listed in Table 1. The log, rather than the native value was used because:

1. The relationship between model outputs and parameter values listed in Table 1 is likely to be more linear with respect to the logs of most of these parameters than with respect to the parameters themselves;
2. Calculation of sensitivities with respect to the logs of parameters involves normalization of sensitivities with respect to current parameter values and therefore, to some extent, with respect to the innate variability of each parameter. Where singular value decomposition is used as a calibration device (as occurs implicitly in computation of the two statistics discussed herein), estimates of scaled parameters are of minimized error variance.

Computation of the parameter error variance reduction statistic requires that the user supply an estimate of the pre-calibration error variance (which is equivalent to the uncertainty) of all model parameters. In the present case this was obtained by first assigning upper and lower bounds to the log of each parameter based on knowledge of the study area. The difference between the two was then divided by four (to compute an approximation to standard deviation) and then squared (to compute variance). This approximation becomes exact if model parameters exhibit independent log-normal prior probability distributions, and if bounds on these parameter define the 95% confidence range of the log of each.

A total of 487 (log-transformed) daily flow measurements comprised the calibration dataset; a uniform weighting was applied. Sensitivities of corresponding model outputs to model parameters were calculated by PEST using finite differences with a 1% parameter perturbation. The boundary between the solution and null subspaces was set at a specific singular value calculated using the PEST SUPCALC utility (Doherty, 2008). The singular value was chosen such that the error variance associated with estimation of combinations of parameters corresponding to additional singular values increases rather than decreases as the solution space is expanded (by attempting estimation of further parameter combinations). Or to put it another way, singular value truncation took place where the error variance accompanying estimation of the eigenvector associated with a particular singular value increases rather than decreases when the additional eigencomponent is included in the solution space. (Note that singular values are arranged in order of decreasing magnitude when implementing this procedure.) It should be emphasized that the number of parameters for which high identifiability is computed is directly related to the number of singular values chosen for the solution-space cut-off. If the solution space cut-off was selected at 18 singular values, an identifiability of 1.0 would be calculated for all

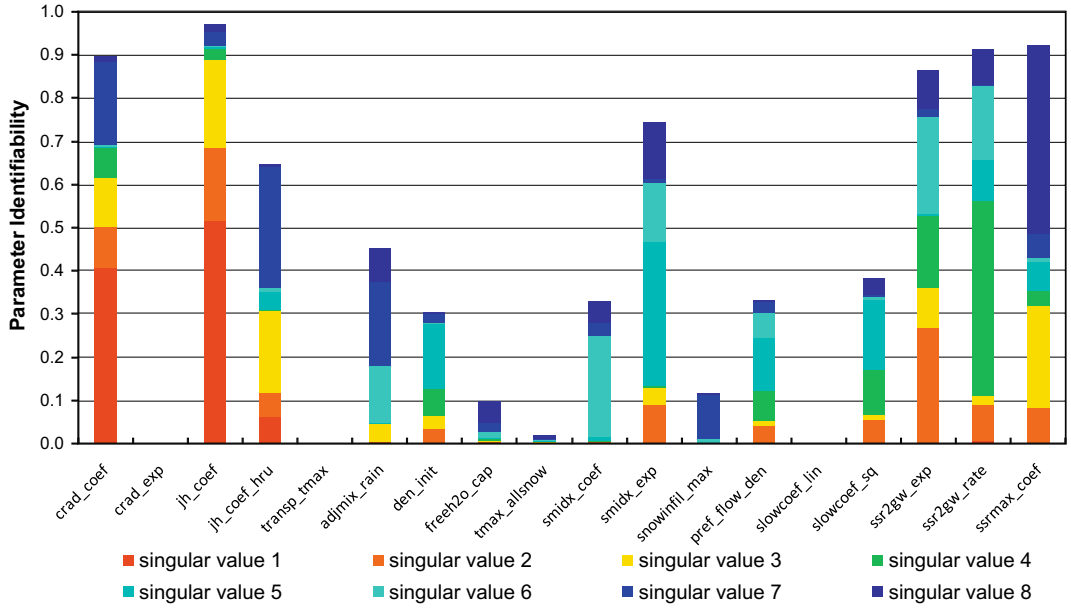

Fig. 3. GSFLOW model parameter identifiability based on a hydrograph of 487 log-transformed daily streamflow values. "Hotter" colors represent singular values of higher magnitude (and lower index); "cooler" colors represent singular values of lower magnitude (and higher index).

parameters. However concomitant calculation of negative values for relative parameter error reduction would reveal the inappropriateness of this choice.

Using results from the SUPCALC utility, eight singular values were assigned to the solution space in the example problem. The IDENTPAR and GENLINPRED utilities (Doherty, 2008) were then employed for calculation of the identifiability and relative error reduction of each parameter listed in Table 1. Parameter identifiabilities are plotted in Fig. 3. The total height of each bar in this figure is the identifiability of the pertinent parameter. Each bar is color-coded according to the contributions made to this identifiability by different eigencomponents spanning the calibration solution space. Hotter colors correspond to eigencomponents with singular values of higher magnitude (i.e., singular values of lower index), while cooler colors correspond to eigencomponents with singular values of lower magnitude (i.e., singular values of higher index).

Parameters featured in Fig. 3 show a large range of identifiability, with some being zero and others approaching one. If a qualitative identifiability level of 0.8 is (somewhat arbitrarily) chosen to mark the cut-off between parameters which are classed as being "identifiable" and those which are classed as not, then five of the 18 sur-

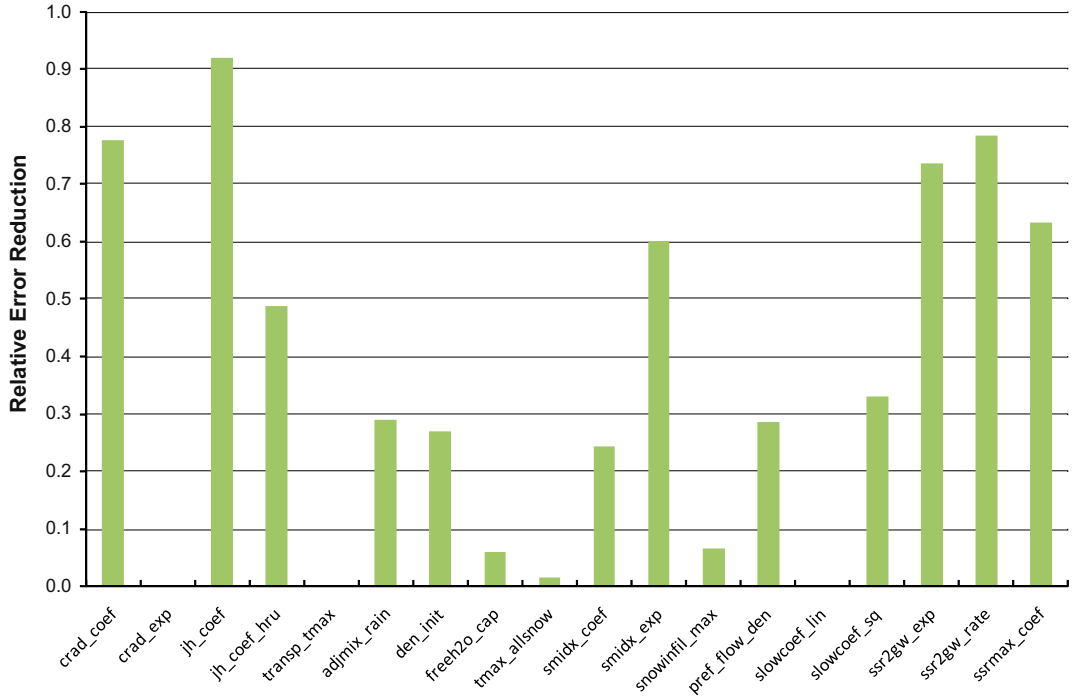

Fig. 4. Relative error reduction of GSFLOW model parameters based on a hydrograph of 487 log-transformed daily streamflow values (parameters are described in Table 1).

face-water model parameters are identifiable on the basis of the single log-transformed daily flow time series – a finding consistent with the conclusions of Bevin (1989) and Jakeman and Hornberger (1993) discussed above. For comparison, analysis of composite parameter sensitivities alone (which does not account for parameter correlation) suggests that all 18 parameters are estimable.

Relative parameter error reductions are plotted in Fig. 4. The correspondence between these and identifiability is good. However, as expected given the increasing error associated with estimation of eigencomponents associated with singular values of decreasing magnitude, differences between the heights of the error reduction bars (Fig. 4) and identifiability bars (Fig. 3) are greater for parameters with identifiability bars dominated by cooler colors in Fig. 3 than for those with bars dominated by warmer colors. The lower magnitudes of singular values indicated by these cooler colors promote greater contribution to parameter estimation error by noise associated with the observation dataset in accordance with the second term of Eq. (A9). If measurement-related noise is statistically time-independent, calibration against a longer flow time series would decrease the contribution of this noise term to the total error variance of the pertinent parameters, as the averaging inherent in model calibration would then tend to "cancel out" the noise. Where noise associated with the dataset is in fact structural noise (and therefore exhibits a temporal and event-based correlation structure), this may not be the case. A more complete discussion of the effect of errors induced by model inadequacy (i.e., structural noise) is beyond the scope of this paper. However, the statistics have identified parameters whose estimated values are likely to be most affected by noise of any kind associated with model outputs and/or their measured counterparts.

Discussion and conclusions

The two statistics discussed in this paper have three primary uses:

1. if computed prior to calibration, they can assist a modeler in deciding which parameters should be included in the calibration process and which should be assigned user-specified values or tied to other parameters;
2. if computed after calibration, they provide a qualitative assessment of the integrity of individual parameter values achieved through the calibration process;
3. they can be employed to test the efficacy of gathering data that are not yet part of the calibration dataset.

The statistics described herein provide a modeler with valuable information on the state of the inverse problem of model calibration as currently formulated. If the problem is ill-posed, they allow the modeler to identify which parameters are responsible. The statistics can also be useful in studies that attempt to "regionalize" parameters employed by models of similar types in different watersheds, in that defensible post-calibration parameter regionalization should be based only on parameters that are identifiable. Another possible use of these statistics is in evaluating potential returns on investment in a complex model to address one or more issues that are faced in a particular study area. A more complex model may indeed provide better simulation of environmental behavior than a simpler one, but this is more likely to hold true if its parameters can all be assigned values that are representative of the area simulated. If the number of identifiable parameters employed by a complex model is no greater than that employed by a simple model, the case for a complex over a simple model becomes weaker.

It is therefore hoped that through use of the statistics discussed here, modelers will gain greater insight into what model calibration, and therefore often modeling itself, can and cannot achieve in a given context. It is further hoped that figures such as Fig. 3, if routinely provided with modeling reports, will provide readers of those reports with a more easily understood representation of the utility of a particular model than is often provided at present.

Software

Software for computing the statistics discussed herein is available free of charge through PEST and its associated utility suite at: http://www.pesthomepage.org/.

Acknowledgements

Support was provided by the USGS Ground-Water Resources Program and the USGS Trout Lake Water, Energy, and Biogeochemical Budgets (WEBB) project. John F. Walker (USGS) is thanked for his help constructing the North Creek PEST input dataset and for review of the draft manuscript. Hoshin Gupta (University of Arizona), Michael Fienen (USGS), and an anonymous reviewer are also thanked for their review of the manuscript. Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the US Government.

Appendix

General

Let $\mathbf{X}$ represent the action of a model on its parameters $\mathbf{p}$. If the model is linear, then $\mathbf{X}$ represents the model itself; if the model is non-linear, then it represents the sensitivity of model outputs for which there are corresponding field measurements to parameters employed by the model. Let $\mathbf{h}$ represent field observations of system state. Then:

$$\mathbf{h} = \mathbf{X}\mathbf{p} + \boldsymbol{\varepsilon} \tag{A1}$$

where $\boldsymbol{\varepsilon}$ represents measurement (and structural) noise. Suppose that there are more elements of $\mathbf{p}$ than can be uniquely estimated. It follows that the matrix $\mathbf{X}$ is column-rank-deficient. Therefore, parameter space can be subdivided into two orthogonal subspaces, a solution subspace and a null subspace. These are spanned by orthogonal unit vectors comprising the columns of two matrices $\mathbf{V}_1$ and $\mathbf{V}_2$. These matrices can be computed through singular value decomposition (SVD) of the matrix $\mathbf{Q}^{1/2}\mathbf{X}$, where $\mathbf{Q}$ is an appropriately chosen observation weighting matrix. Thus:

$$\mathbf{Q}^{1/2}\mathbf{X} = [\mathbf{U}_1 \quad \mathbf{U}_2] \begin{bmatrix} \mathbf{S}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_2 \end{bmatrix} \begin{bmatrix} \mathbf{V}_1^t \\ \mathbf{V}_2^t \end{bmatrix} \tag{A2}$$

In Eq. (A2) the superscript "t" represents matrix transpose. The columns of the $\mathbf{U}_1$ and $\mathbf{U}_2$ matrices are unit vectors that span the range space of $\mathbf{X}$. $\mathbf{S}_1$ and $\mathbf{S}_2$ are diagonal matrices containing singular values of $\mathbf{Q}^{1/2}\mathbf{X}$, higher-valued ones in $\mathbf{S}_1$ and lower-valued and zero-valued ones in $\mathbf{S}_2$. The cut-off between the calibration solution and null spaces is chosen on the basis of separation of these two sets of singular values. Moore and Doherty (2005) show that a suitable truncation point (i.e. the point at which higher singular values are assigned to $\mathbf{S}_1$ and lower singular values are assigned to $\mathbf{S}_2$) can be formulated as that for which the error variance of a particular prediction of interest is minimized. Doherty (2008) suggests choosing the truncation point at that singular value at which the error variance of a "prediction" whose sensitivities to parameters are provided by the elements of the complimentary eigenvector (column of $\mathbf{V}$) is raised rather than lowered through its inclusion in the solution space rather than in the null space.

Parameter. identifiability

The "identifiability" of parameter *i*, and its complement the "non-identifiability" of parameter *i*, are defined as:

$$f_i = (\mathbf{V}_1\mathbf{V}_1^t)_{ii} = \mathbf{i}^t(\mathbf{V}_1\mathbf{V}_1^t)\mathbf{i} \tag{A3a}$$

and

$$g_i = (\mathbf{V}_2\mathbf{V}_2^t)_{ii} = \mathbf{i}^t\mathbf{V}_2\mathbf{V}_2^t\mathbf{i} \tag{A3b}$$

where the subscript "*i*,*i*" indicates the *i*th diagonal element of a matrix, and the vector **i** is a unit vector in which all elements except that pertaining to the parameter in question are zero. It is easily shown that:

$$f_i + g_i = 1 \tag{A4}$$

for all parameters p_i. The following properties of f_i and g_i are worth noting:

1. if a vector of unit length pointing in the direction of parameter p_i is projected onto the calibration solution space, f_i is the length of this projected vector. It follows from the definition of the inner product that f_i is also the cosine of the angle between the unit vector and its projection;
2. if a vector of unit length pointing in the direction of parameter p_i is projected onto the calibration null space, g_i is the length of this projected vector. It is also, again, the cosine of the angle between the unit vector and its projection;
3. f_i is the *i*th diagonal element of the so-called "resolution matrix", a much-used concept in regularization theory. As is explained in texts such as Aster et al (2004) and Menke (1984), the closer the resolution matrix is to the identity matrix, the greater the extent to which solution of the inverse problem allows "perfect resolution" of estimated parameters. Where the inverse problem is ill-posed, it is mathematically impossible for the resolution matrix to equal the identity matrix.

Parameter error

Let *s* be a prediction made by a model and let the sensitivity of that prediction to all parameters be encapsulated in the vector **y**. Thus, for a linear model:

$$s = \mathbf{y}^t\mathbf{p} \tag{A5}$$

Moore and Doherty (2005) show that the post-calibration error variance associated with a prediction is:

$$\sigma_s^2 = \mathbf{y}^t(\mathbf{I}-\mathbf{R})C(\mathbf{p})(\mathbf{I}-\mathbf{R})^t\mathbf{y} + \mathbf{y}^t\mathbf{G}C(\boldsymbol{\varepsilon})\mathbf{G}^t\mathbf{y} \tag{A6}$$

where **R** (the resolution matrix) and **G** are matrices that depend on the method chosen for solution of the ill-posed inverse problem that is model calibration. C(**p**) is a covariance matrix that describes the pre-calibration uncertainty of parameters (and can include any natural correlation between them) while $C(\boldsymbol{\varepsilon})$ is the covariance matrix of noise associated with the measurement dataset. If truncated SVD is used as a regularization device then:

$$\mathbf{I}-\mathbf{R} = \mathbf{V}_2\mathbf{V}_2^t = \mathbf{I}-\mathbf{V}_1\mathbf{V}_1^t \tag{A7a}$$

and

$$\mathbf{G} = (\mathbf{V}_1\mathbf{S}^{-1}\mathbf{V}_1)\mathbf{X}^t\mathbf{Q} \tag{A7b}$$

Suppose that the weight matrix **Q** is chosen such that it is proportional to the inverse of measurement noise (as is common practice). Thus:

$$C(\boldsymbol{\varepsilon}) = \sigma_r^2\mathbf{Q}^{-1} \tag{A8}$$

where σ_r^2 is a so-called "reference variance". Suppose further that the prediction is in fact the value of parameter p_i, so that σ_s^2 becomes the post-calibration error variance of that parameter and **y** becomes **i**, a unit vector in the direction of the parameter. Thus the post-calibration error variance of parameter *i*, which we denote as $[\sigma_2^2]_i$, becomes:

$$[\sigma_2^2]_i = \mathbf{i}^t\mathbf{V}_2\mathbf{V}_2^tC(\mathbf{p})\mathbf{V}_2\mathbf{V}_2^t\mathbf{i} + \sigma_r^2\mathbf{i}^t\mathbf{V}_1\mathbf{S}_1^{-1}\mathbf{V}_1^t\mathbf{i} \tag{A9}$$

Pre-calibration error variance, denoted herein as $[\sigma_1^2]_i$, is calculated as:

$$[\sigma_1^2]_i = \mathbf{i}^tC(\mathbf{p})\mathbf{i} \tag{A10}$$

This follows from (A9); alternatively, it can be recognized as simply the *i*'th diagonal element of C(**p**).

We define the "relative parameter error" e_i of parameter *i* as its post-calibration error variance divided by its pre-calibration error variance. That is:

$$e_i = [\sigma_2^2]_i/[\sigma_1^2]_i \tag{A11a}$$

Its complement, the "relative error reduction" of parameter *i* is defined as:

$$r_i = 1 - e_i \tag{A11b}$$

r_i has a maximum value of one. Unfortunately its minimum value can be less than zero; this can occur if too many dimensions are assigned to the calibration solution space, and measurement errors thus contaminate parameter estimates to too great an extent. The reason for this phenomenon is the fact that solution space singular values appear in the denominator (rather than the numerator) of the second term of Eq. (A9); the occurrence of singular values in $\mathbf{S}_1$ that are too small can lead to unduly large elements of $\mathbf{S}_1^{-1}$, thus "amplifying" the contribution that measurement noise makes to $[\sigma_2^2]_i$. This is the reason why "overfitting" during calibration can lead to estimates for parameter values that can actually amplify potential model predictive error beyond its pre-calibration level.

References

Aster, R., Borchers, B., Thurber, C., 2004. Parameter Estimation and Inverse Problems. Elsevier Academic Press. p. 301.

Bastidas, L.A., Gupta, H.V., Sorooshian, S., Shuttleworth, W.J., Yang, Z.L., 1999. Sensitivity analysis of a land surface scheme using multicriteria methods. Journal of Geophysical Research-Atmospheres 104 (D16), 19481–19490.

Beck, M.B., Halfon, E., 1991. Uncertainty, identifiability and the propagation of prediction errors: a case study of Lake Ontario. Journal of Forecasting 10, 135–161.

Beven, K.J., Binley, A.M., 1992. The future of distributed models: model calibration and uncertainty prediction. Hydrological Processes 6, 279–298.

Beven, K.J., Freer, J., 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modeling of complex environmental systems. Journal of Hydrology 249 (1–4), 11–29.

Bevin, K.J., 1989. Changing ideas in hydrology – the case of physically based models. Journal of Hydrology 105, 157–172.

Bicknell, B.R., Imhoff, J.C., Kittle, J.L., Jobes, T.H., Donigian, A.S., 2001. HSPF User's Manual. Aqua Terra Consultants, Mountain View, California.

Boyle, D.P., Gupta, H.V., Sorooshian, S., 2000. Toward improved calibration of hydrologic models: combing the strengths of manual and automatic methods. Water Resources Research 36 (12), 3663–3674.

Burnash, R.J.C., Ferral, R.L., McGuire, R.A., 1973. A Generalized Streamflow Simulation System – Conceptual Modeling for Digital Computers. US Department of Commerce, National Weather Service and State of California, Department of Water Resources.

Doherty, J., 2008. Manual and Addendum for PEST: Model Independent Parameter Estimation. Watermark Numerical Computing, Brisbane, Australia.

Doherty, J., Johnston, J.M., 2003. Methodologies for calibration and predictive analysis of a watershed model. Journal American Water Resources Association 39 (2), 251–265.

Doherty, J., Skahill, B., 2006. An advanced regularization methodology for use in watershed model calibration. Journal of Hydrology 327 (3–4), 564–577.

Harbaugh, A.W., 2005. MODFLOW-2005, the US Geological Survey modular groundwater model-the Ground-Water Flow Process: US Geological Survey Techniques and Methods, vol. 6-A16 (variously paginated).

Hill, M.C., Tiedeman, C.R., 2007. Effective Groundwater Model Calibration: With Analysis of Data, Sensitivities, Predictions and Uncertainty. Wiley and Sons. p. 464.

Hunt, R.J., Doherty, J., 2006. A strategy for constructing models to minimize prediction uncertainty. In: MODFLOW and More 2006 – Managing Ground Water Systems: Proceedings of the Seventh International Conference of the International Ground Water Modeling Center. Colorado School of Mines, Golden, CO, pp. 56–60.

Hunt, R.J., Feinstein, D.T., Pint, C.D., Anderson, M.P., 2006. The importance of diverse data types to calibrate a watershed model of the Trout Lake Basin, northern Wisconsin, USA. Journal of Hydrology 321 (1–4), 286–296.

Hunt, R.J., Doherty, J., Tonkin, M.J., 2007. Are models too simple? Arguments for increased parameterization. Ground Water 45 (3), 254–262. doi:10.1111/j.1745-6584.2007.00316.x.

Hunt, R.J., Walker, J.F., and Doherty, J., 2008a. Using GSFLOW to simulate climate change in a northern temperate climate. In: MODFLOW and More 2008: Ground Water and Public Policy, Proceedings of the Ninth International Conference of the International Ground Water Modeling Center. Colorado School of Mines, Golden, CO, pp. 109–113.

Hunt, R.J., Prudic, D.E., Walker, J.F., Anderson, M.P., 2008b. Importance of unsaturated zone flow for simulating recharge in a humid climate. Ground Water 46 (4), 551–560. doi:10.1111/j.1745-6584.2007.00427.x.

Jakeman, A.J., Hornberger, G.M., 1993. How much complexity is warranted in a rainfall–runoff model? Water Resources Research 29 (8), 2637–2649.

Leavesley, G.H., Lichty, R.W., Troutman, B.M., Saindon, L.G., 1983. Precipitation–runoff modeling system – user's manual: US Geological Survey Water-Resources Investigations Report 83-4238, 207 p.

Marcé, R., Ruiz, C.E., Armengol, J., 2008. Using spatially distributed parameters and multi-response objective functions to solve parameterization of complex applications of semi-distributed hydrological models. Water Resources Research 44, W02436. doi:10.1029/2006WR005785.

Markstrom, S.L., Niswonger, R.G., Regan, R.S., Prudic, D.E., Barlow, P.M., 2008. GSFLOW—Coupled Ground-Water and Surface-Water Flow Model Based on the Integration of the Precipitation–Runoff Modeling System (PRMS) and the Modular Ground-Water Flow Model (MODFLOW-2005). Techniques and Methods, vol. 6–D1. 240 p. <http://pubs.usgs.gov/tm/tm6d1/>.

Menke, W., 1984. Geophysical Data Analysis: Discrete Inverse Theory. Academic Press Inc.

Moore, C.M., Doherty, J., 2005. The role of the calibration process in reducing model predictive error. Water Resources Research 41 (5), W05050. doi:10.1029/2004WR003501.

Pint, C.D., 2002. A Groundwater Flow Model of the Trout Lake Basin, Wisconsin: Calibration and Lake Capture Zone Analysis. M.S. thesis, Department of Geology and Geophysics, University of Wisconsin-Madison.

Poeter, E.P., Hill, M.C., Banta, E.R., Mehl, S., Christensen, S., 2005. UCODE_2005 and Six Other Computer Codes for Universal Sensitivity Analysis, Calibration, and Uncertainty Evaluation. US Geological Survey Techniques and Methods, vol. 6-A11, 283 p.

Sorooshian, S., Gupta, V.K., 1983. Automatic calibration of conceptual rainfall–runoff models: the question of parameter observability and uniqueness. Water Resources Research 19 (1), 260–268.

Tonkin, M.J., Tiedeman, C.R., Ely, D.M., Hill, M.C., 2007. OPR-PPR, a computer program for assessing data importance to model predictions using linear statistics. US Geological Survey Techniques and Methods Report, Book 6, 115 p (Chapter E2).

Vrugt, J.A., Bouten, W., Gupta, H.V., Sorooshian, S., 2002. Towards improved identifiability of hydrologic model parameters: the information content of experimental data. Water Resources Research 38 (12), 1312. doi:10.1029/2001WR001118.

Wagener, T., McIntyre, N., Lees, M.J., Wheater, H.S., Gupta, H.V., 2003. Towards reduced uncertainty in conceptual rainfall–runoff modelling: dynamic identifiability analysis. Hydrological Processes 17 (2), 455–476.

Walker, J.F., Bullen, T.D., 2000. Trout Lake, Wisconsin – A Water, Energy, and Biogeochemical Budgets program site. USGS Fact Sheet 161-99. <http://water.usgs.gov/pubs/fs/fs-161-99/pdf/fs-161-99.pdf>.

Yeh, W.W.-G., 1986. Review of parameter identification procedures in groundwater hydrology: the inverse problem. Water Resources Research 22 (2), 95–108.