Detection of Overparameterization and Overfitting in an Automatic Calibration of SWAT

G. Whittaker, R. Confesor, Jr., M. Di Luzio, J. G. Arnold

Abstract. *Distributed hydrologic models based on small-scale physical processes tend to have a large number of parameters to represent spatial heterogeneity. This characteristic requires the use of a large number of parameters in model calibration. It is a common view that calibration with a large number parameters produces overparameterization and overfitting. Recent work using prior information, spatial information, and constraints on parameters for regularization of the calibration problem has improved model predictions using a few dozen parameters. We demonstrate that the Soil and Water Assessment Tool (SWAT) and the information associated with a SWAT watershed setup provide a regularized problem with many of recently published regularization techniques already utilized in SWAT. Our hypothesis is that the Soil and Water Assessment Tool (SWAT) regularizes the inverse problem so that a stable solution can be obtained for calibration of SWAT using a very large number of parameters, where very large means up to 10,000 calibration parameters. In this study, a two-objective calibration genetic algorithm based on a non-dominated sorting genetic algorithm (NSGA-II) was used to calibrate the Blue River basin in Oklahoma. We introduce the use of intermediate solutions found by the genetic algorithm to test identification of calibration parameters and diagnose model overfitting. Defining identification as the capability of a model to constrain the estimation of parameters, we introduced a method for statistically testing for changes from the initial uniform distribution of each parameter. We found that all 4,198 parameters used to calculate the Blue River SWAT model were identified. Diagnostic comparisons of goodness-of-fit measures for the calibration and validation periods provided strong evidence that the model was not overfitted.*

Keywords. *Automatic calibration, Distributed hydrologic model, NSGA-II, Overfitting, Overparameterization, Regularization.*

Distributed hydrologic models typically use parameters that are not directly observed and must be calibrated to the observed characteristics of a watershed (Beven, 2000; Reed et al., 2004). Hydrological model calibration in this way is categorized as an inverse problem, since the parameters that are used to predict streamflow (and other parameters) are chosen using the observed streamflow. In some sense, calibration works backwards from observation so that observed and predicted output values are in agreement. As the hydrologic model increases in complexity, the number of parameters in the model may increase to the point where overparameterization can result in overfitting and deterioration in prediction accuracy (Jakeman and Hornberger, 1993; Beven, 1993, 2006).

Approaches to hydrologic prediction have been categorized in two ways: (1) an upward or mechanistic approach, in which small-scale physical processes and landscape heterogeneity are modeled as much as possible, and (2) a downward or statistical approach (Schoups et al., 2008; Wagener et al., 2007). When small-scale physical processes are modeled throughout a large, hetereogeneous landscape, the number of parameters can become very large. For example, the Soil and Water Assessment Tool (SWAT; Arnold et al., 1998) used in this study is designed for simulation of agricultural practices, fate and transport of soil and chemicals, and basin hydrology. A study watershed is divided into subwatersheds, where each subwatershed is characterized by hydrologic response units (HRU). Each HRU is a unique combination of landuse/landcover, soil, and slope. SWAT is categorized as a semi-distributed model because the output of each HRU is placed at the mouth of the subwatershed that contains the HRU. By choosing a small value for subbasin size, high-resolution simulations are possible, as well as the standard USGS HUC, or other divisions of the landscape. Given the broad scope of the physical processes that are included in SWAT, there are several hundred parameters set by default for every HRU. In this study, we used the Blue River basin in Oklahoma, which is set up in SWAT with 55 subbasins and 193 HRUs. There are over 4,000 total parameters to be calibrated, even if only 21 parameters in each HRU are selected for calibration.

Submitted for review in November 2009 as manuscript number SW 8317; approved for publication by the Soil & Water Division of ASABE in August 2010.

The authors are **Gerald Whittaker,** Research Hydrologist, USDA-ARS National Forage Seed Production Research Center, Corvallis, Oregon; **Remegio B. Confesor, Jr., ASABE Member,** Senior Research Scientist, National Center for Water Quality Research, Heidelberg University, Tiffin, Ohio; **Mauro Di Luzio,** Research Scientist, Texas AgriLife Research, Blackland Research Center,Texas A&M University System, Temple, Texas; and **Jeffrey G. Arnold, ASABE Member Engineer,** Supervisory Agriculture Engineer, USDA-ARS Grassland Soil and Water Research Laboratory, Temple, Texas. **Corresponding author:** Gerald Whittaker, USDA-ARS, 3450 SW Campus Way, Corvallis, OR 97331; phone: 541-738-4157; fax: 541-738-4160; e-mail: Jerry.whittaker@ars.usda.gov.

The simulation of diverse regions also forces the use of a large number of parameters. The SWAT literature database for peer-reviewed journal articles (https://www.card.iastate.edu/swat_articles/) has links to SWAT applications (569 articles as of 30 October 2009) on all continents, and includes most developed countries. Given this range of application, SWAT must have sufficient model complexity to represent different hydroclimatic regimes worldwide (van Werkhoven et al., 2009).

Overparameterization is defined in statistics as the inclusion of redundant information, and the effect is to produce a singular covariance matrix that cannot be inverted. This is a serious problem for calibration methods that require matrix inversion, but it is no obstacle for genetic algorithms. For application of the concept to hydrologic inverse modeling, overparamerization has been defined as the situation in which the amount of information contained in a single hydrograph used for calibration is not enough to estimate a large number of parameters (Jakeman and Hornberger, 1993). This definition is a special case of the statistical definition.

An overparameterized calibration can produce an overfitted model. Overfitting refers to a circumstance in which a model starts fitting the noise in the calibration data set, as well as the calibration observations. The result of an overfitted model is a reduced error for the calibration data set, but larger errors in validation and prediction with other data sets. For example, if a model has been overfit to a five-year hydrograph for 1995-2000, the predictions of the model for other time periods will be worse than would be expected based on the goodness-of-fit of the calibration. While increasing the number of parameters that are used in a calibration makes it more likely that overfitting will occur, note that this definition is independent of the number of parameters.

The issue of overparameterization and overfitting must be resolved before meaningful analysis of sensitivity and uncertainty can be accomplished. If a model is overparameterized and/or overfitted, the parameter estimates and error estimates will be unreliable. A calibrated model that is not overparameterized or overfitted is a necessary condition for sensitivity and uncertainty analysis.

Tikhonov regularization (Tikhonov and Arsenin, 1977), a method of adding linear constraints to a problem, has been used in hydrological studies with several variations to deal with overparameterization and overfitting. The addition of constraints allows the inclusion of prior information about the parameters to be used in calibration, including equality constraints and limits on the range of parameters. Constraints can alleviate overparameterization by constraining the feasible region of parameters. For example, in the absence of regularization, if two correlated parameters have the same range of values, there will almost certainly be overparameterization. If the parameters are constrained by regularization to different ranges, overparameterization disappears. Constraints on parameters and their interaction by regularization can in effect create a filter that only responds to the signal (hydrograph) and ignores the noise, eliminating overfitting. Useful extensions of the regularization method include addition of linear constraints where the parameters are weighted in comparison with observed parameters (Doherty and Skahill, 2006; Tonkin and Doherty, 2005) and spatial regularization to constrain the feasible parameter space (Pokhrel and Gupta, 2009; Pokhrel et al., 2008; Pokhrel et al., 2009). Model coupling provides another method of regularization by adding information from other physical processes that constrain simulation of hydrologic processes. Hinnell et al. (2009) find that independent geophysical properties in a geophysical model coupled to hydrologic properties in an inverse model can provide reductions in errors in hydrologic predictions.

We observe that SWAT includes all of these methods of regularization: equality constraints, inequality constraints, constraints on parameter bounds, prior information on a wide range of geophysical data, all data spatially referenced, and independent physical processes linked to hydrology. Some examples of the regularization properties of SWAT are (1) coupled biological models, e.g., a plant growth model relating removal of water and nutrients from the root zone, transpiration, and biomass/yield production; (2) coupled biophysical models, e.g., the nitrogen cycle; (3) physical models, e.g., evapotranspiration; (4) prior information, e.g., georeferenced temperature, rainfall, and soil characteristics; (5) equality constraints on parameters, e.g., several hundred parameters are set for each HRU by default in SWAT or with prior information and will not be used in a particular calibration; and (6) parameter bounds constraints, e.g., most parameters in the model are limited to a specific range taken from the literature, and it is common procedure in calibration to set a parameter range acceptable to the analyst for the calibration parameters.

The general objective of this article is to analyze the effect of regularization provided by the SWAT model on calibration with a large number of parameters. The specific objectives are: (1) to test for the occurrence of overparameterization after automatic calibration of SWAT, and (2) to determine whether the resulting SWAT models calibrated with a large number of parameters are overfitted. We present the methods used to achieve these two objectives, and the results and statistical inferences associated for each of the objectives. If a model is overparameterized and/or overfitted, the results of sensitivity and uncertainty analysis will almost certainly be invalid, and the analyst should turn to parameter reduction or other regularization techniques. The analysis in this study is only the first step in the evaluation of an automatic calibration. This study does not include sensitivity analysis, parameter selection, and uncertainty analysis since these concepts are not necessary in testing for overfitting and overparameterization.

In the next section, we state our hypothesis and contribution to the literature. The study area is described in the following section. The multiobjective automatic calibration method using a genetic algorithm is summarized in the section titled "Multiple-Objective SWAT Calibration Approach." The statistical test for identification and over-parameterization that we developed for genetic algorithms is explained and applied in the section titled "Nonparametric Test for Identification." Summary statistics of model prediction error are compared in the "Diagnosis of Overfitting" section, and the results are used to draw inferences about the hypothesis in the conclusion.

Hypothesis and Experimental Design

Our hypothesis is that the Soil and Water Assessment Tool (SWAT) regularizes the inverse problem so that a stable solution can be obtained for calibration of SWAT using a very large number of parameters when set up with spatially referenced information for the simulation watershed.

Our primary contribution to the literature is the evaluation of the above hypothesis. As discussed earlier, it is not generally expected that inclusion of a large number of parameters in calibration will result in a simulation model with good forecasting capabilities. If it can be shown that SWAT is adequately calibrated even with the use of a large number of parameters, then automatic calibration is rather simple to implement, and it becomes an attractive alternative to the time-consuming and computationally expensive process of parameter reduction. If the hypothesis is accepted, one could expect a calibration to be more representative of similar watersheds where data were not available, since it is less likely that a necessary parameter has been omitted in the process of parameter reduction.

To test our hypothesis, we statistically analyze the intermediate solutions and the time path of the optimization algorithm. To our knowledge, this is the first time that this information has been used in a study of hydrologic models, and we could only find one other study that looked at the information produced during the optimization search (Marseguerra et al., 2003). In the analysis of the intermediate data produced by the optimization algorithm, we demonstrate a novel statistical procedure for testing for overparameterization, and we introduce a simple qualitative method of testing for model overfitting.

Blue River Study Watershed

We selected the Blue River watershed located in southern Oklahoma. It is narrowly extended with a draining area of around 1233 km^2 and is characterized by a gentle topography (fig. 1). The elevation ranges from around 427 m to about 154 m at U.S. Geological Survey (USGS) gauging station 7332500 at the outlet of the watershed near the city of Blue. The soils of the watershed are characterized by very high volumetric clay content (Smith et al., 2004). Analysis of the digital landuse/landcover data set used in this study yielded a dominant distribution of natural or semi-natural (35%) and planted or intensively managed herbaceous vegetation (32%), deciduous forest (23%), evergreen forest (3%), cropland (2%), and irrelevant portions of urban and rangeland areas.

Smith et al. (2004) report a mean annual precipitation of 1036 mm and a runoff coefficient, estimated from annual data, of 0.17. The Blue River watershed is a particularly challenging case study. The narrow shape of the watershed imposes precise precipitation field records. Note that the hydrologic behavior in the watershed area is affected by a high hydrogeologic complexity (Smith et al., 2006). This influence appears particularly remarkable at the upstream reaches, where USGS discharge station 7332390 near Connerville is located (fig. 1).

Simulation Configuration

We used the SWAT model version 2005. The model is extensively illustrated in Neitsch et al. (2005). We set the model with the Green and Ampt infiltration method with an hourly simulation time step (Di Luzio and Arnold, 2004a). We used a GIS software tool to configure the input data for AVSWAT-X (Di Luzio and Arnold, 2004b; Di Luzio et al., 2004), to delineate streams and subbasins, and to define the HRUs required by the model simulation. We used the following digital data sets: 30 m National Elevation Dataset DEM (Gesch, 2007; Gesch et al., 2002), 30 m USGS National Land Cover Data (Vogelmann et al., 2001), and STATSGO (State Soil Geographic Database) 1:250,000-scale soil map layer (USDA, 1992). Using the interface, we identified and established a topography-driven connection of 55 subbasins containing 193 HRUs. Proper vegetation and soil hydraulic parameters were extracted from a database included in the model and from the soil parameters estimated using the MUUF (Map Unit User Files) method (Baumer et al., 1994) respectively.

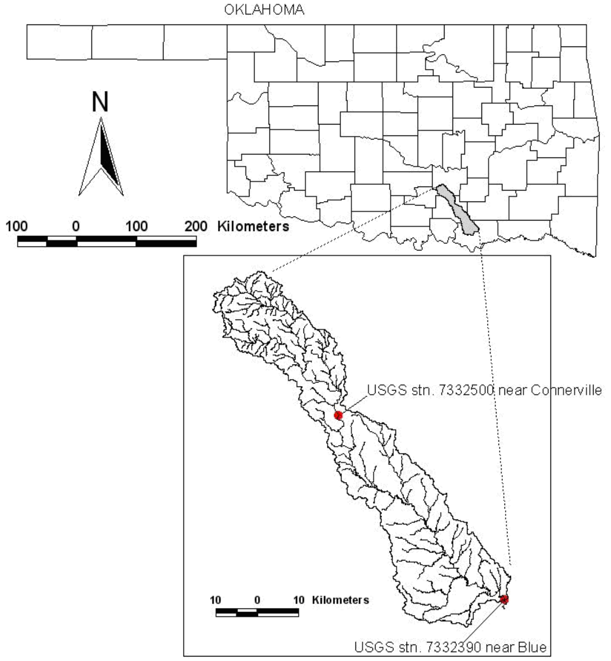
OKLAHOMA
N
100 0 100 200 Kilometers
USGS stn. 7332500 near Connerville
10 0 10 Kilometers
USGS stn. 7332390 near Blue

Figure 1. Location of the Blue River watershed. The two USGS discharge stations in the watershed are depicted.

Climatic input data included hourly precipitation records at each subbasin. We calculated the records as mean areal values for each subbasin using the hourly 4 km × 4 km radar grids taken from the Distributed Model Intercomparison Project (DMIP-2, initiated by the National Weather Service) (Smith et al., 2006) and the respective HRAP (Hydrologic Rainfall Analysis Project; Reed and Maidment, 1999) coordinate system. Daily air temperature data were associated to the closest measuring gauge part of the NCDC Cooperate Observer Stations network.

Multiple-Objective SWAT Calibration Approach

We used the method of automatic calibration described by Confesor and Whittaker (2007). The method is an implementation of the nondominated sorting genetic algorithm (NSGA-II, Deb et al., 2002a), a fast and efficient multiple objective optimizing algorithm characterized by a nondominated sorting algorithm, an elitist selection method, and the elimination of a sharing parameter. NSGA-II assigns fitness by Pareto ranking (or nondomination) and crowding distance to the combined parent and child populations. The solution is then ranked by domination, where a solution X1 dominates another solution X2 if two conditions are satisfied (Deb, 2001): (1) the solution X1 is no worse than X2 in all objectives, and (2) the solution X1 is strictly better than X2 in at least one objective. Crowding distance is the average distance between an individual and its nearest neighbors in the search space. In minimization problems, each solution is ranked by the number of solutions that dominate it. Those that are dominated by a large number of solutions will have a lower fitness and a lower chance of surviving in the population. In cases where the solutions have the same nondomination rank, the solution with larger crowding distance is preferred, ensuring a diverse population spread across the Pareto optimal front.

The genetic algorithm was set up with a population of 100 individuals, where each individual was a set of calibration parameters. Each individual consisted of 4,198 values, the total number of parameters used in calibrating the model. The genetic algorithm was real-coded to simplify parameter representation. We found that a population of 100 converged as fast as using 200 individuals. In a genetic algorithm, mutation is the random substitution of values into the genome, and cross-over is the mathematical representation of sexual reproduction. For real-valued genomes (many genetic algorithms use binary values in the genome), various schemes of linear combination of genes from two parents are used as a cross-over operator. In our study, the probability of mutation was 0.1, and we used the PCX cross-over operator with a level of 3 (Deb et al., 2002b). The algorithm is programmed in the R statistical language (R Development Core Team, 2007). Our previous simulation runs showed that the calibration stabilized after 250 generations characterized by no apparent change in the Pareto optimal front. However, in this study, we stopped after 1,000 generations as added insurance that no significantly better solutions could be found.

Table 1. Parameters used in calibration of the SWAT model of the Blue River, initialization range and geographic resolution.

Parameter	Definition	Limits[a]	Resolution[b]
PHU	Heat units to bring plant to maturity	0.0 - 1500	HRU
SOL_K	Saturated hydraulic conductivity (mm/h)	0.75 - 1.1*	HRU_S
SOL_AWC	Available water capacity (mm/mm)	0.75 - 1.1*	HRU_S
SOL_CRK	Crack volume potential	0.0 - 0.3	HRU
CH_N2	Manning's n for main channel	0.014 - 0.075	Sub
OV_N	Manning's n for overland flow	0.1 - 4.0	HRU
CANMX	Maximum canopy storage (mm)	0.0 - 5.0	HRU
ESCO	Soil evaporation compensation factor	0.1 - 1.0	HRU and W
EPCO	Plant water uptake comp. factor	0.1 - 1.0	HRU and W
REVAPMN	Threshold depth (mm)	0.001 - 200.0	HRU
ALPHA_BF	Baseflow alpha factor (days)	0.04 - 1.0	HRU
GW_DELAY	Groundwater delay time (days)	0.0 - 60.0	HRU
GW_REVAP	Groundwater revap coefficient (days)	0.02 - 0.20	HRU
SURLAG	Surface runoff lag coefficient	1.0 - 21.0	W
MSK_CO1	Calibration coefficient	0.0 - 3.0	W
MSK_CO2	Calibration coefficient	0.0 - 5.0	W
MSK_X	Weighting factor	0.0 - 0.50	W
TRNSRCH	Transmission loss	0.10 - 0.90	W
EVRCH	Reach evaporation adjustment factor	0.10 - 0.90	W
SLSUBBSN	Average slope length (m)	0.75 - 1.25*	HRU
SLSOIL	Slope length for lateral slope length (m)	0.0 - 30.0	HRU
HRU_SLP	Average slope steepness (m/m)	0.75 - 1.25*	HRU
TIMP	Snow pack temperature lag factor	0.01 - 1.00	W
SMFMN	Minimum melt factor for snow (mm/°C day)	1.4 - 6.9	W
SMFMX	Minimum melt factor for snow (mm/°C day)	1.4 - 6.9	W
CHL_1	Longest tributary channel length in subbasin	0.75 - 1.25*	SUB
CH_S1	Average slope of tributary channels	0.75 - 1.25*	SUB
CH_W1	Average width of tributary channels	0.75 - 1.25*	SUB
CH_N1	Manning's n for the tributary channels	0.75 - 1.25*	SUB
CH_K1	Effective hydraulic conductivity of tributary	0.025 - 10.000	SUB
CH_L2	Length of main channel	0.75 - 1.25*	SUB
CH_S2	Average slope of main channel	0.75 - 1.25*	SUB
CH_W2	Average width of main channel	0.75 - 1.25*	SUB
CH_D	Average depth of main channel	0.75 - 1.25*	SUB
CH_K2	Effective hydraulic conductivity of main channel	0.025 - 10.000	SUB
CH_WDR	Channel width to depth ratio	0.75 - 1.25*	SUB
ALPHA_BNK	Alpha factor for bank storage recession curve	0.001 - 0.990	SUB
GWQMN	Threshold depth of water in shallow aquifer	0.000 - 200.0	HRU
RCHRG_DP	Recharge to deep aquifer	0.0 - 1.0	HRU
GW_SPYLD	Specific yield for shallow aquifer	0.001 - 0.009	HRU

[a] Parameters marked with an (*) are multipliers of the default value in the SWAT model setup.

[b] W = watershed, Sub = subbasin, HRU = HRU, and HRU_S = HRU soil layer.

For application to automatic calibration of SWAT, we used parameters selected by experts. Some of the parameters were multipliers of the default values, and others were the actual value used in SWAT. Where multipliers are used, the parameter values are ensured to be within the reasonable range derived from the geographic information on soils and land use used in the initial setup of the model using the ArcGIS interface. The range of the remaining parameters is restricted to values considered reasonable in the opinion of the authors. The population of the genetic algorithm was initialized by random draws from the ranges given in table 1. Two objectives were defined for the evaluation, based on an automatic baseflow filter (Arnold et al., 1998) separation of two components of flow, i.e., event driven and base flow. The streamflow was designated as driven when the automatic filter first-pass base flow was <80% of the observed streamflow; otherwise, the streamflow was classified as nondriven (Boyle et al., 2001). Minimizing the root mean square error (RMSE) of the predicted base flow was defined as one objective function. The second objective function was minimizing the RMSE of the predicted event driven flow. RMSE was defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Q_{obs,i} - Q_{sim,i})^2} \qquad (1)$$

where n is the number of time steps with peak or low flow events, $Q_{obs,i}$ is the observed streamflow at time i, and $Q_{sim,i}$ is the simulated streamflow at time i. We are aware that other approaches to model calibration offer advantages when analysis of sensitivity and model error diagnosis are important (Vrugt et al., 2009). However, our objective for this study was to provide forecasts with no error terms or other related statistical analysis. This is a common situation in application of hydrologic models, especially in consulting.

For the production of a single forecast, an individual from the Pareto optimal front is selected based on the preferences of the analyst. Where there is no reason to prefer one objective over another, we choose an individual that weights the two objectives approximately equally.

Nonparametric Test for Identification

Identification is the capability of the model to constrain the parameters used in the model (Doherty and Hunt, 2009). Following a suggestion by Marseguerra et al. (2003), we saved the values of the population and their evaluations at each generation of the genetic algorithm. Marseguerra et al. (2003) demonstrated that the evolution of a parameter in a genetic algorithm indicates sensitivity by the narrowing of the variance of the population of values of that parameter, and the order in which the change of variance occurs. We extend their work with the observation that a reduction in variation for the population of a parameter means that the genetic algorithm has constrained the range of estimated parameters. This is in effect the definition of identification that Doherty and Hunt (2009) suggested. This concept is illustrated in figure 2, where one parameter in the model is identified (dashed line) and shows considerable reduction in variance in the population. The parameter that is not identified (solid line) shows little change in the variance of the parameter value in the population.

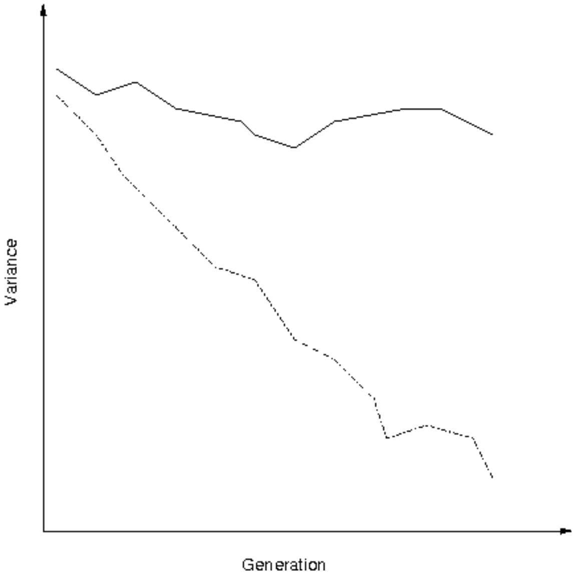

Figure 2. Comparison of change in the variance of two parameters' values with differing sensitivity during calibration using a genetic algorithm.

We apply a nonparametric test of equality of distribution combined with discriminant analysis to construct a novel procedure for statistically testing parameters for identification. We claim that the parameter is identified when the distribution of parameter estimates is significantly different from the uniform distribution used for the initialization of all parameters.

The basis for our proposed test identification is the fact that different combinations of parameter values will result in different probabilities of being retained in the genome due to selection at the evaluation stage of the genetic algorithm. Where a parameter has no effect on the fitness of genome, there is said to be no selection pressure on that parameter. That is, no matter what the value of the parameter, the probability of survival of the individual in the population is not affected. In a genetic algorithm, the lack of selection pressure results in an unconstrained parameter, by our definition a parameter that is not identified. To illustrate the distribution of a parameter that is not identified, i.e., a parameter with no selection pressure, we added a dummy parameter to the genome of the genetic algorithm. This dummy parameter had no effect on the evaluation of the objectives. Nonparametric density estimation using the averaged shifted histogram (ASH; Scott, 1992) provides a convenient method to visualize the change in distribution of a parameter during the evolution of the optimal solution set. Figure 3 shows the distribution of this dummy parameter with a population of 100 through 500 generations, where the parameter was randomly initialized between 0 and 1500. Note that the distribution, with minor variations, remains constant and approximately uniform throughout the optimization. The effect of identification on the distribution of a parameter is obvious in figure 4, where a uniform distribution was quickly constrained at about generation 50, and shifted and narrowed even more at generations 250 to 350. Since the genetic algorithm used for calibration was initialized with a uniform distribution of parameter values in the feasible space, and we know that without selection pressure the distribution will not change, any observed change in the distribution of values of parameter indicates that the parameter has been identified

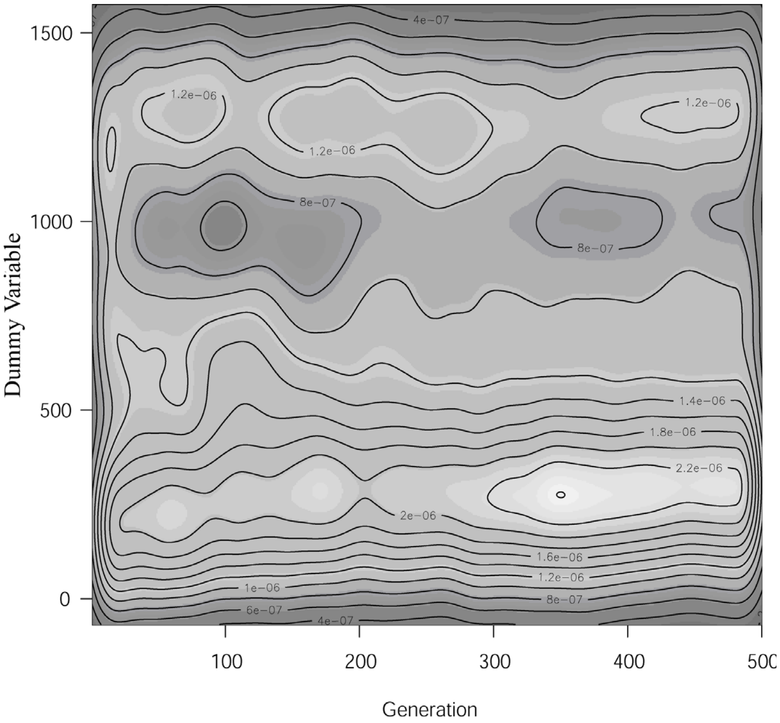

Figure 3. Kernel density estimate of probability distribution function of population of a dummy parameter through 500 generations.

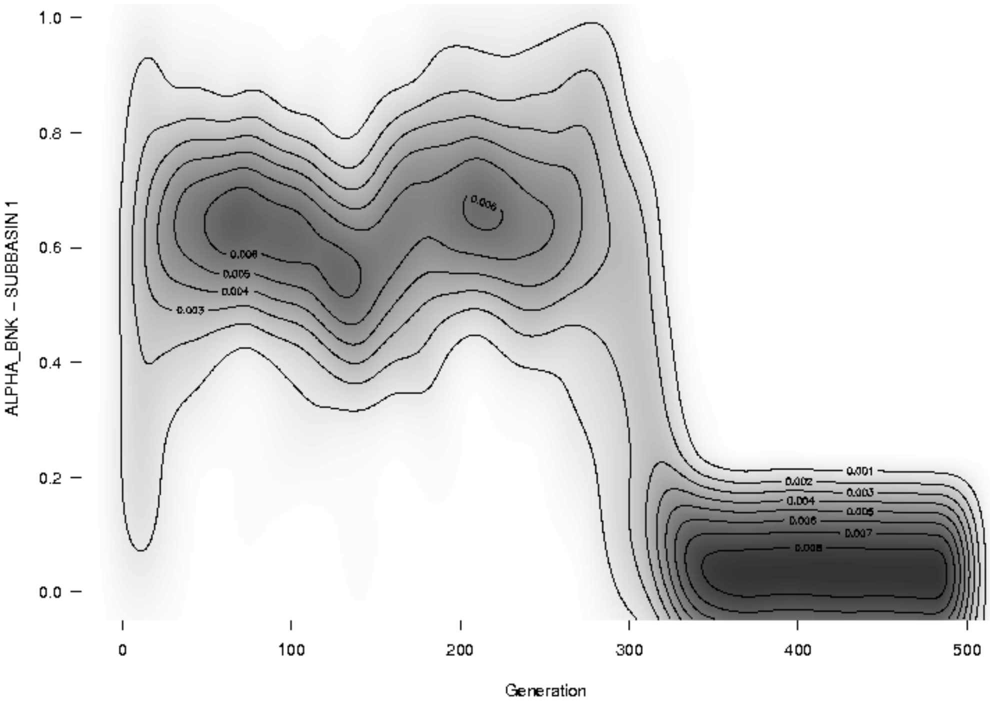

Figure 4. Kernel density estimate of probability distribution function of population of ALPHA_BNK in Subbasin 1 through 500 generations.

(i.e., constrained by the model during the optimization). We take advantage of this fact to statistically test all parameters for identification.

The procedure that we followed for this test is:

1. Save the calibration parameters and simulated flows for each individual in the population at each generation.
2. Choose a generation to serve as baseline for intergenerational comparison of the distribution of values of each calibration parameter.
3. Calculate the stochastic equality hypothesis test statistic between the baseline generation and all other generations, for each parameter.
4. Calculate the time path for the stochastic equality hypothesis test statistic for each parameter.

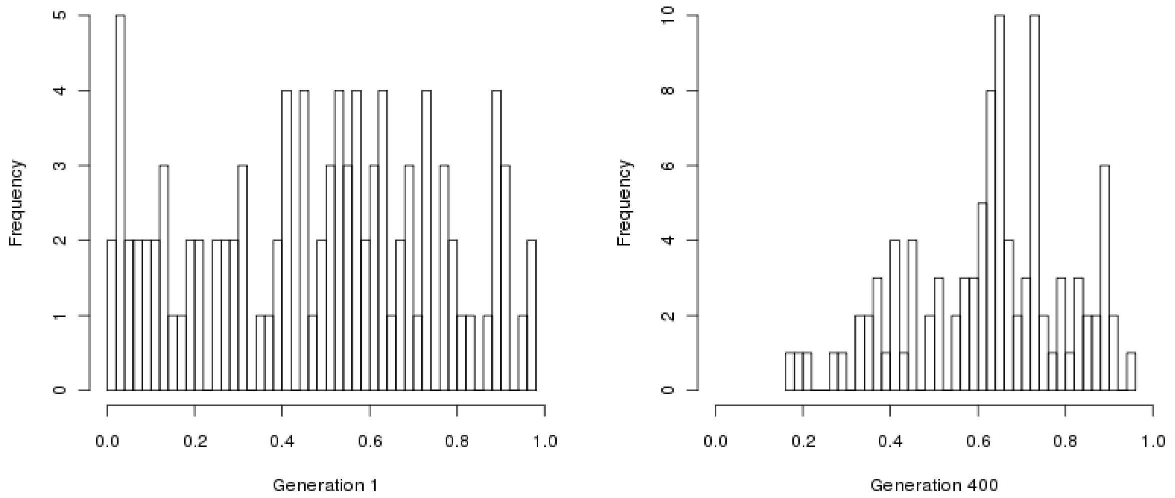

Figure 5. Comparison of dispersal of population values for alpha_bnk (HRU 10) in the initial draw and after 400 generations.

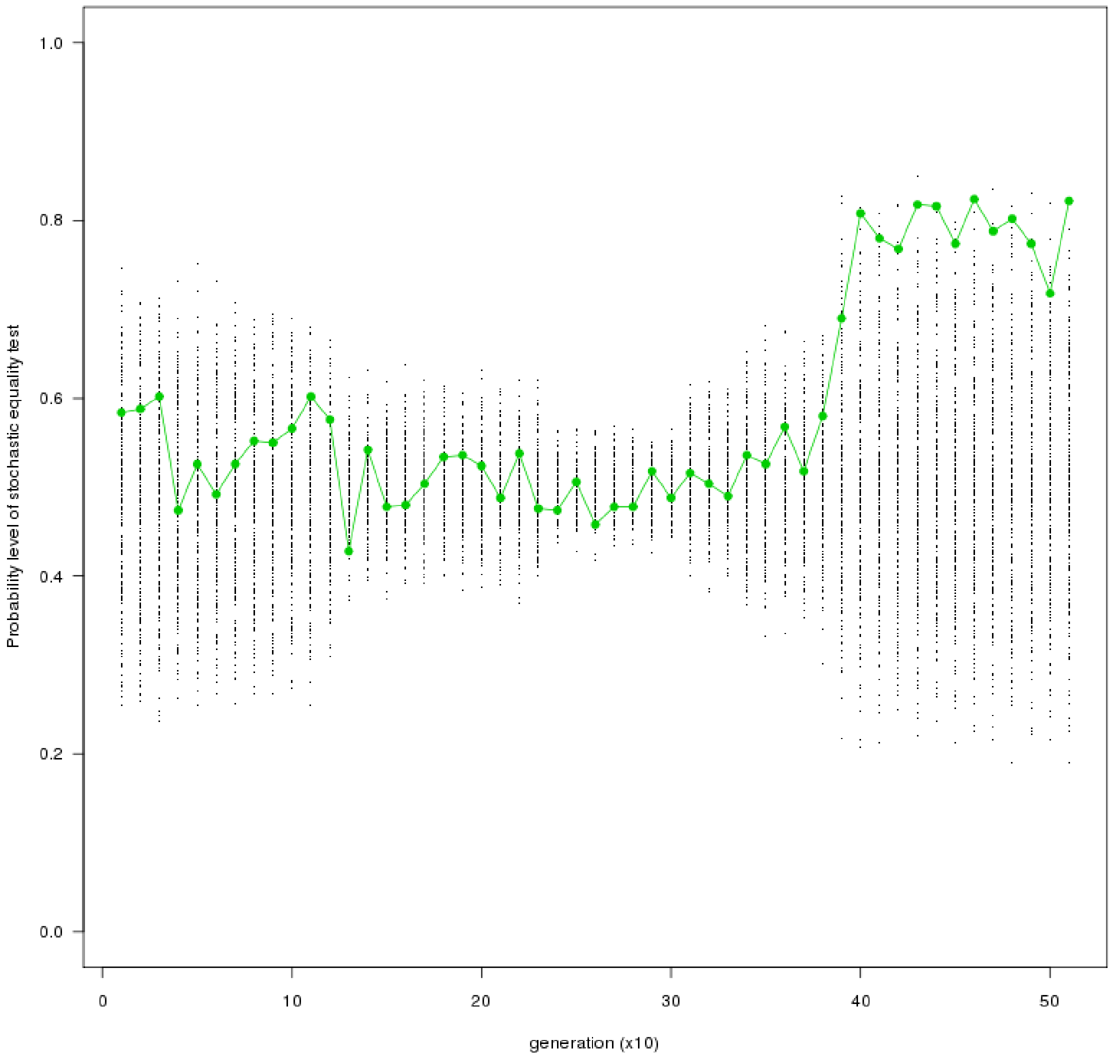

Figure 6. Time path for dummy parameter 562 (no effect on model output) of probability of stochastic equality of every tenth generation with generation 250.

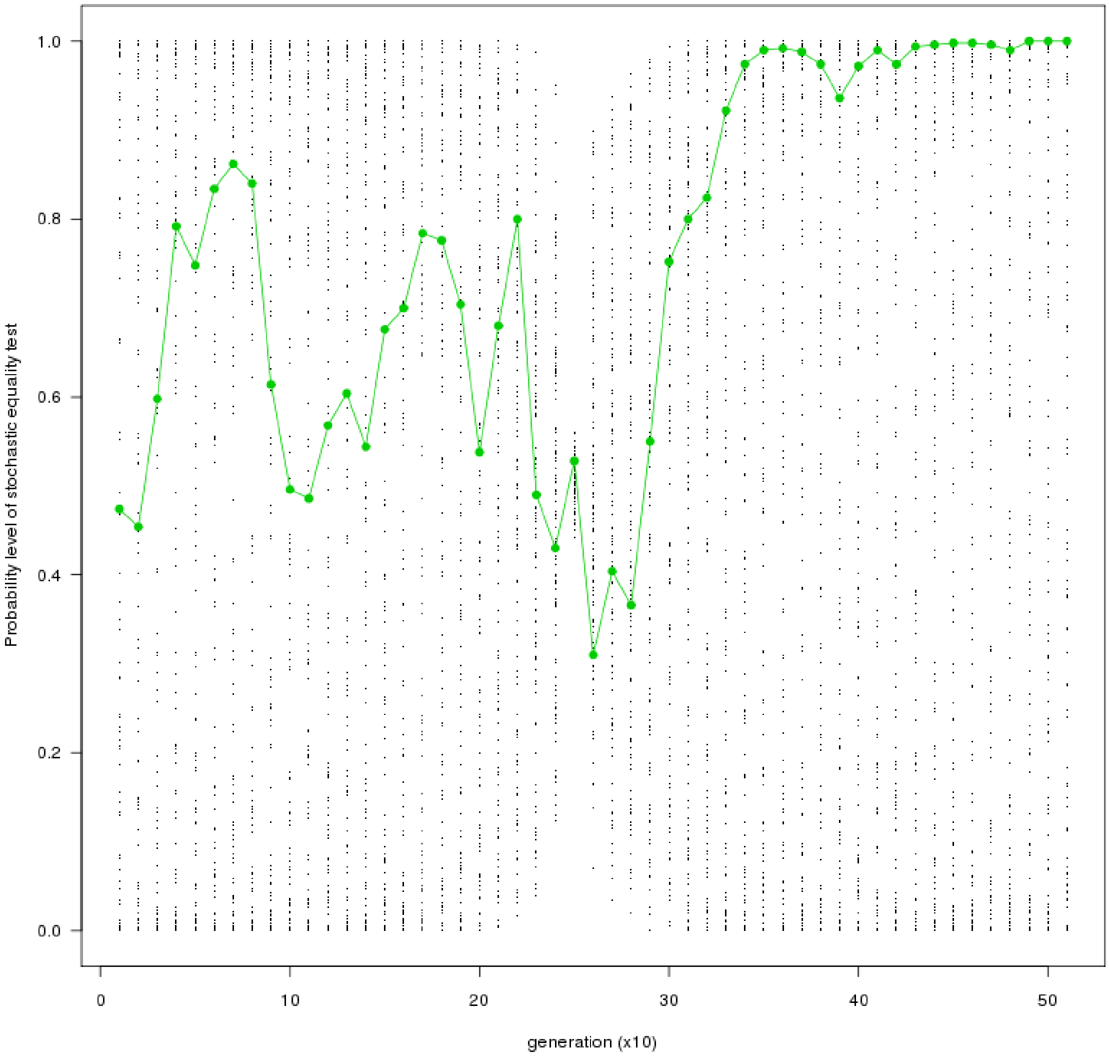

Figure 7. Time path of probability of stochastic equality of every tenth generation with generation 250, alpha_bnk, HRU 10.

5. Evaluate four descriptive measures of the time path of the test statistic for each parameter.
6. Apply linear discriminant analysis to identify constrained and unconstrained parameters.
7. If all parameters are constrained, then the model is not overparameterized.

From cross-sections of the distributions shown in figures 3 and 4, it is easy to see that the distribution in a given generation is not normal, and the use of a measure of variance based on the assumption of normality is inappropriate. Figure 5 shows the non-normality of the distribution of values of an identified parameter at different generations. Therefore, a test for a change in distribution between generations cannot be based on assumption of normality or other distributions. Reiczigel et al. (2005) introduced a test of stochastic equality of two non-normal populations based on the bootstrap. Stochastic equality is based on the probability that two non-normal distributions are the same. The function is available in R code at: www.univet.hu/users/jreiczig/brw/ (accessed 11 November 2009).

To statistically test for a change in a parameter distribution during optimization, we ran the method of Reiczigel et al. (2005) to test for stochastic equality. Tests of stochastic equality calculate the probability that two non-normal distributions are the same. More formally, the hypothesis test is H_0: $P(X<Y) = P(X>Y)$ against $P(X<Y) \neq P(X>Y)$, where the equality and inequality relations are called stochastic equality and stochastic inequality. The test statistic is based on the rank Welch statistic (t_{rw}). We applied the test to compare the distribution of parameters in generation 250 with all others in the first 500 generations using a thousand bootstrap samples for each test. To establish a baseline for comparison, we ran the calibration algorithm for 500 generations with 4,198 dummy parameters, i.e., no SWAT parameters were used. The hypothesis test results for a population size of 100 in each generation with 4,198 parameters for each individual are displayed in figure 6. Each generation has the test statistic of 4,198 comparisons of stochastic equality with that of generation 250. The choice of generation 250 was arbitrary, and any other generation should lead to the same conclusions. The y-axis shows the probability that the distribution of each parameter differs from the distribution of that parameter in generation 250. A small number of parameters approach the 90% confidence limit. At the 5% confidence limit for both tails, where the change in the distribution is significant, there are no dummy parameters. The pattern that we expect to see for parameters that are not identified is represented by the path followed by dummy parameter 562.

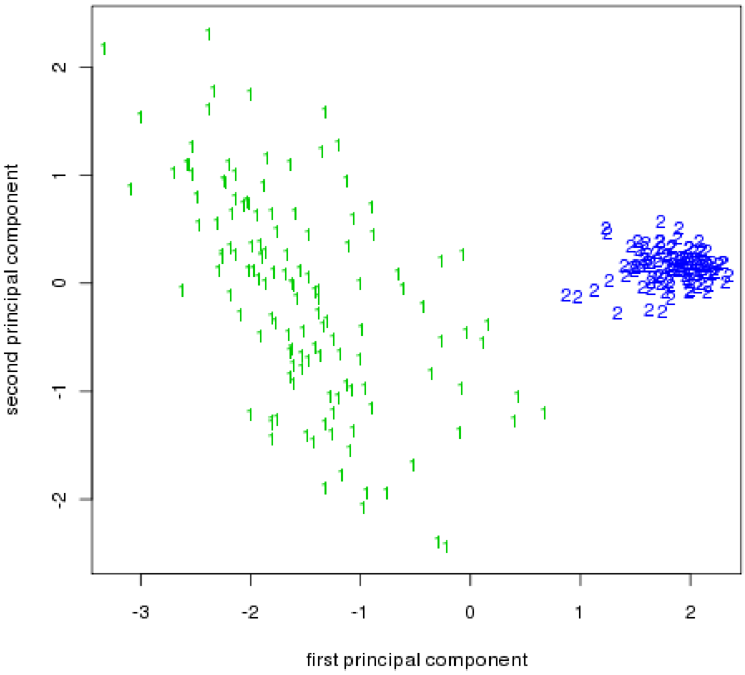

Figure 8. Principal components analysis of training set used in screening with linear discriminant analysis. Synthetic unconstrained parameters are represented by "1", and parameters selected randomly from the calibration data set are represented by "2".

The test statistic shows that the distribution of each parameter changes with generational remove, but in 250 generations does not reach 95% level of significance for t_{rw}.

Figure 7 shows the results of the stochastic equality test applied to the evolution of the distributions of the 4,198 parameters in the Blue River calibration. The difference in the general pattern is immediately obvious, with a large number of significantly different distributions. The time pattern of the identified parameter traced in figure 7 differs in its significance level from the time path of the parameter lacking identification highlighted in figure 6. However, we needed an automatic, statistical method rather than visual analysis of 4,198 time paths to finally detect whether each parameter is identified. Based on many visualizations, we chose four summary measures to describe the behavior of the stochastic equality test through time. The measures each show a different aspect of the speed and range of the change in distribution when the genetic algorithm used the parameter in the optimization. The measures are: (1) the sum of the absolute value of changes in t_{rw} ; (2) the range, i.e., the biggest change in t_{rw}; (3) max[abs(min/max)] of t_{rw} ; and (4) the fraction of generations between 1-150 and 300-450, with t_{rw} between 0.2 and 0.8.

To screen all parameters, we applied linear discriminant analysis (LDA). LDA calculates the probability of an observation belonging to a group, where the effect of a parameter is calculated by training the algorithm on a data set with known categories (supervised classification; Venables and Ripley, 2002). To apply LDA, the algorithm is "trained" on a known sample and then applied to the remainder of the data. We created a data set with 4,198 parameters that we knew were not identified by running the calibration algorithm with the evaluation step replaced by a draw from a uniform random distribution. In this case, the parameters could not have any effect on the evolution of the solution and were not correlated with the evaluation. This procedure ensured that the distribution of each parameter in the not-identified data set would have the characteristics of an unconstrained parameter that was only changed through genetic drift. To apply LDA, a training data set with known categories is required, but the categories of the parameters in the calibration data set were unknown. Since the categories in the calibration data were unknown, we used principal component analysis (PCA), an unsupervised classification method, to test if the calibration data were in a second category that could be assumed to be identified. Figure 8 shows that PCA clearly identified two categories in a randomly selected training data set. All of one category is not identified (represented by "1" in fig. 8), and the other category consists exclusively of calibration data sets ("2"). We analyzed a large number of randomly drawn training data sets and could find none that did not show the same pattern. A test calibration run in which ten dummy parameters were included in the genome with 4,198 SWAT parameters gave the same result. The dummy parameters were clearly in a different category from the SWAT parameters. When mixed with synthetic not-identified parameters, the dummy parameters from the calibration were clearly categorized with the not-identified parameters. Based on this evidence, we assumed that the parameters from the calibration in the training data set represented identified parameters for comparison with the not-identified synthetic parameters.

We trained LDA on a random sample of 100 from each of the two data sets and assumed that the unknown data set from the actual calibration was identified and formed a second category. When we applied LDA to the calibration data set, every parameter was classified as identified. We ran the procedure with ten different randomly selected training data sets, with the same results each time. This result is strong evidence that all calibration parameters were identified, following the definition of Doherty and Hunt (2009). A sample of dummy and synthetic not-identified parameters was included as a check on the method, and all were correctly classified in every run.

Diagnosis of Overfitting

Our approach to diagnose overfitting in calibration of a hydrologic model is based on the well-known effect that the predictions of a model usually become worse when the number of parameters is increased (Doherty and Hunt, 2009; Faber and Rajko, 2007). A corollary to this behavior is that when a model is calibrated starting with random parameters, the fit generally improves during the genetic algorithm's search for a Pareto optimal set. During this part of the search, the fit of the predictions will improve for both the calibration and validation data sets. When a model is overparameterized, at some point the calibration will start fitting noise in the calibration data set, and the fit of predictions for the validation data set will degrade while the fit for calibration data set continues to improve. For application of these observations to diagnosis of whether a model is overfitted or not, we compare calibration error measures and validation error measures throughout the fitting of the genetic algorithm to see if degradation of the validation model fit can be observed. By following the change in fit as the genetic algorithm finds solutions with smaller errors, we should be able to detect when the model starts fitting noise.

Several criteria are commonly used in evaluation of the accuracy of a model calibration and validation (Moriasi et al. 2007), including Nash-Sutcliffe efficiency (NSE; Nash and Sutcliffe, 1970), a measure of how close the simulated parameter matches observations of the parameter of interest:

$$\text{NSE} = 1 - \left[\frac{\sum_{i=1}^{n} (Y_i^{obs} - Y_i^{sim})^2}{\sum_{i=1}^{n} (Y_i^{obs} - Y^{mean})^2} \right] \tag{2}$$

where $Y_i{}^{obs}$ is the ith observation of parameter Y, $Y_i{}^{sim}$ is the ith value of the simulated value of Y, and Y^{mean} is the estimated mean of parameter Y. A value of 1 is the best performance, greater than 0 is acceptable, and less than or equal to 0 is not acceptable.

RMSE was used in the objective function, but a related measure is the RMSE normalized by the standard deviation of the measured data, as recommended by Moriasi et al. (2007). This measure is called the RMSE-observations standard deviation ratio (RSR) and ranges from 0 (perfect model simulation) to large positive numbers for a poor simulation. The RSR is useful for comparison of different constituents with different scales and is defined as:

$$\text{RSR} = \frac{\text{RMSE}}{\text{SD}_{obs}} = \frac{\sqrt{\sum_{i=1}^{n} (Y_i^{obs} - Y_i^{sim})^2}}{\sqrt{\sum_{i=1}^{n} (Y_i^{obs} - Y^{mean})^2}} \tag{3}$$

Percent model bias (PBIAS; Gupta et al., 1999) is proportional to the tendency of model to over- or underestimate the parameter of interest:

$$\text{PBIAS} = \left[\frac{\sum_{i=1}^{n} (Y_i^{obs} - Y_i^{sim}) * (100)}{\sum_{i=1}^{n} (Y_i^{obs})} \right] \tag{4}$$

Figures 9, 10, and 11 compare the RSR, NS, and PBIAS measures of the models at selected generations between 1 and 2000, respectively. There are 100 individuals (combinations of parameters used in the calibration) in each generation, meaning that there are 100 calibrated models in each generation that can be evaluated. We used the empirical mode (50th percentile) of the goodness-of-fit measure in each generation as a summary statistic for comparison. In all three figures, the goodness-of-fit measure follows the pattern that would be expected for a model that was not overfitted (Radtke and Wong, 2006). In the early generations, the fit improves very quickly for both the calibration and validation data sets. The normalized root mean square error (RSR) continues to improve for 2000 generations for the calibration, while the validation RSR stabilizes at a value of 0.58 by generation 500. The percent bias measure for validation approaches the best values at around 200 generations, and the Nash-Sutcliffe efficiency for validation stabilizes at 500 generations. None of the three measures shows the increase in errors that is expected if a model is overfitted (Radtke and Wong, 2006).

The explanation for how the RSR measure of prediction error can continue to improve in the calibration period while remaining constant in the validation period can be seen in figure 12, where the predicted flows for 800 hours in the validation period for the best-fitting individuals (using RMSE as criterion) in generations 20 and 2000 are compared with the observed flow in the validation period for 800 hours. The predicted baseflow approaches the observed minimum flows rather better after 2000 generations, but the results are mixed in fitting the event flow peaks. At A, the peak prediction of

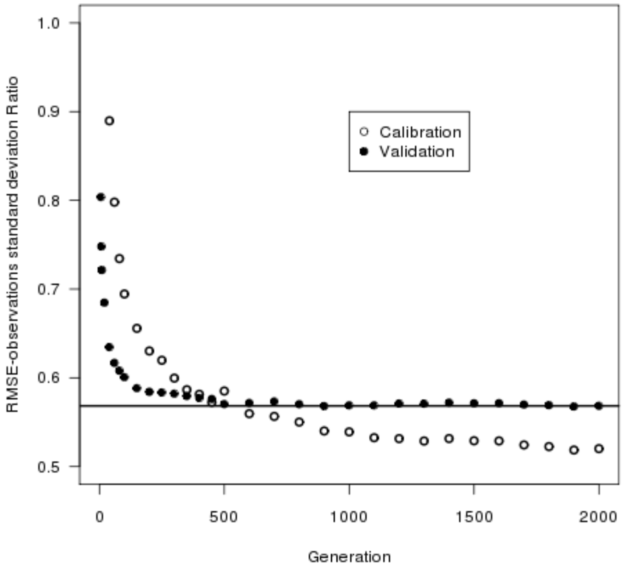

Figure 9. Comparison of calibration and validation goodness-of-fit: mode of normalized root mean square error (RSR) of predictions for the calibration and validation periods using the 100 parameter sets in selected generations of the calibration genetic algorithm. The horizontal line is set at the value of the mode of the predictions for the validation period using the calibration from generation 2000.

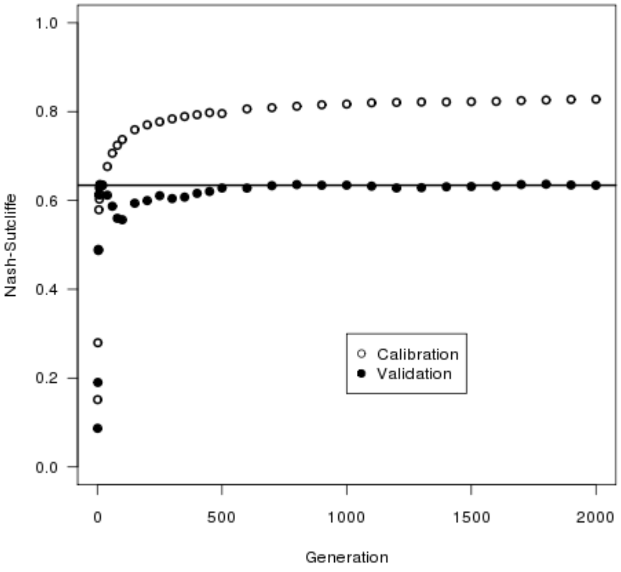

Figure 10. Comparison of calibration and validation goodness-of-fit: mode of Nash-Sutcliffe efficiency (NS) of predictions for the calibration and validation periods using the 100 parameter sets in selected generations of the calibration genetic algorithm. The horizontal line is set at the value of the mode of the predictions for the validation period using the calibration from generation 2000.

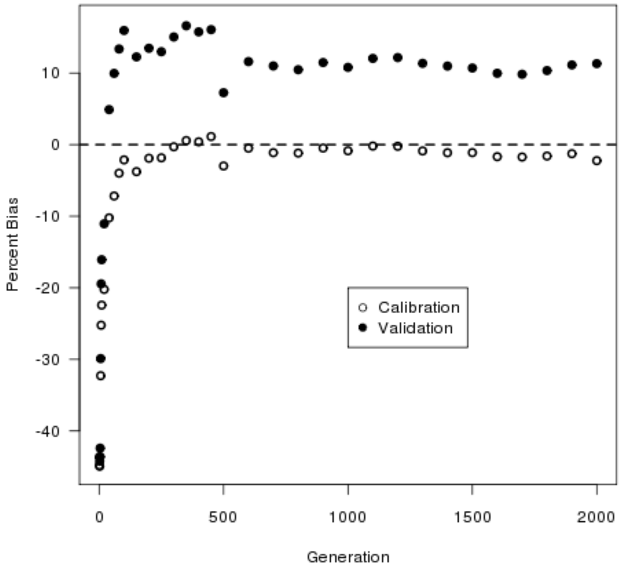

Figure 11. Comparison of calibration and validation goodness-of-fit: mode of percent bias (PBIAS) of predictions for the calibration and validation periods using the 100 parameter sets in selected generations of the calibration genetic algorithm. The horizontal line is set at the value of the mode of the predictions for the validation period using the calibration from generation 2000.

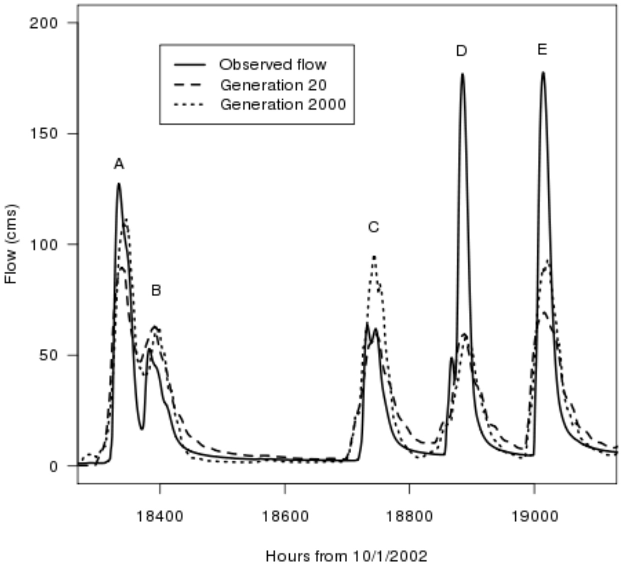

Figure 12. Comparison of observed and predicted flows in a portion of the validation period using parameters from the individuals with the lowest RMSE in generations 20 and 2000.

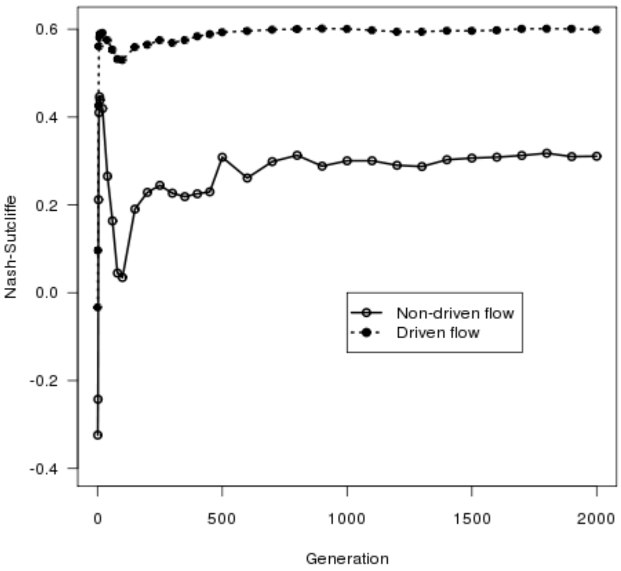

Figure 13. Comparison of driven and non-driven flow goodness-of-fit: mode of Nash-Sutcliffe efficiency (NS) of predictions for the calibration and validation periods using the 100 parameter sets in selected generations of the calibration genetic algorithm.

the younger model is poorer but in the right hour, while the model at 2000 generations has come closer to the magnitude of the flow but is offset two hours from observed maximum. At B, the younger calibration overshot the observed peak and was offset late; calibration to 2000 generations did not really improve the prediction. At C, the prediction from the model in generation 2 is very good but deteriorates badly by generation 2000. At D, SWAT underestimates the maximum, and calibration does not improve the fit. The comparison of fits at E is similar to A, with the fit improving with more generations. These results show that the error of predictions in the validation period may improve or deteriorate in different parts of the hydrograph as the genetic algorithm continues, but the overall effects on the RSR measure cancel so that the measure remains approximately constant after a certain number of generations. The NS measures show this behavior only slightly, while the PBIAS measures do not show it at all. It is also noteworthy that the calibration also yielded good NS and PBIAS of the simulated results from the nested inner subwatershed near Connerville, indicating that the method is predicting a stable subbasin and does not distort the spatial distribution of parameter values as well as the spatial distribution of runoff and perhaps constituent concentration.

It has been reported that the use of RMSE for a calibration criterion is sensitive to peak flows and biases simulations in the recession period (non-driven flow) of the hydrograph (Boyle et al., 2000), resulting in lower efficiency for the non-driven flow. We observe this effect in figure 13, where the Nash-Sutcliffe efficiency of the non-driven flow in the validation period for this study is less than the driven. Apparently, the use of two objectives, one for the driven flow and one for the non-driven flow, did not overcome this bias.

We define a model as overfitted when the model starts to predict the variation from the observed value at each time step, i.e., the "noise" in a general (not statistical) use of that term. It seems extremely unlikely that SWAT prediction of noise in one time series would be generalized enough to predict the noise in a different time series. This is particularly unlikely where rainfall and temperature are important drivers. However, there is nothing that says that this is impossible. This seems like another point that supports the use of a statistical estimator for the model prediction errors, rather than the simple difference of predictions and observations that is widely used in hydrologic model calibration. Where the structure of the errors is modeled, it may be possible to evaluate more precisely if, and how, the noise is fitted by the model. This is the next step in our research in this area.

Conclusion

We tentatively accept the hypothesis that the Soil and Water Assessment Tool (SWAT) regularizes the inverse problem so that a multiobjective calibration of SWAT using 4,198 parameters is not overparameterized or overfitted in the application to the Blue River, Oklahoma, data set.

To investigate the success of our approach, we analyzed the evolution of the calibration parameters. The solutions found by the genetic algorithm at each generation of the calibration provided information to test the identification of calibration parameters and diagnose model overfitting. Defining identification as the capability of a model to constrain the estimation of parameters, we introduced a method for statistically testing for changes from the initial uniform distribution of each parameter. We found that all parameters had statistically different distributions after calibration and concluded that all parameters were identified. Comparison of goodness-of-fit measures of predicted values for the calibration and validation periods showed that because of regularization of the calibration, SWAT was not calibrated to model the noise in the calibration hydrograph and therefore was not overfitted by calibration with 4,198 parameters.

The statistical methods that we used in testing for overparameterization and overfitting are well known, but the application of those methods to information provided by the evolutionary path of the calibration parameters is novel. Our study was limited in that we did not address the issue of parameter sensitivity and uncertainty. We intend to expand our work using the information in the evolutionary time path of a genetic algorithm calibration optimization to investigate parameter sensitivity.

Our general conclusion is that calibration of large, complex physical models may not always require parameter reduction to avoid overparameterization and overfitting. The size and complexity of these models provide constraints that regularize the calibration problem when many parameters are used in a calibration.

References

Arnold, J. G., R. Srinivasan, R. S. Muttiah, and J. R. Williams. 1998. Large-area hydrologic modeling and assessment part I: Model development. *J. American Water Resources Assoc.* 34(1): 73.

Baumer, O., P. Kenyon, and J. Bettis. 1994. *MUUF v2.14 User's Manual*. Lincoln, Neb.: USDA-NRCS National Soil Survey Center.

Beven, K. 1993. Prophecy, reality, and uncertainty in distributed hydrological modeling. *Advances in Water Resources* 16(1): 41-51.

Beven, K. 2000. *Rainfall-Runoff Modeling: The Primer.* New York, N.Y.: John Wiley and Sons.

Beven, K. 2006. A manifesto for the equifinality thesis. *J. Hydrol.* 320(1-2): 18-36.

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

Boyle, D. P., H. V. Gupta, S. Sorooshian, V. Koren, Z. Zhang, and M. Smith. 2001. Toward improved streamflow forecasts: Value of semidistributed modeling. *Water Resources Res.* 37(11): 2749-2759.

Confesor, R. B., and G. Whittaker. 2007. Automatic calibration of hydrologic models with multi-objective evolutionary algorithm and Pareto optimization. *J. American Water Resources Assoc.* 43(4): 1:9.

Deb, K. 2001. *Multi-Objective Optimization Using Evolutionary Algorithms.* Chichester, U.K.: John Wiley and Sons.

Deb, K., A. Pratap, S. Agarwal, and T. Meyarivan. 2002a. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Computation* 6(2): 182-197.

Deb, K., A. Anand, and D. Joshi. 2002b. A computationally efficient evolutionary algorithm for real-parameter optimization. *Evol. Comput.* 10(4): 371-395.

Di Luzio, M., and J. G. Arnold. 2004a. Formulation of a hybrid calibration approach for a physically based distributed model with NEXRAD data input. *J. Hydrol.* 298: 136-154.

Di Luzio, M., and J. G. Arnold. 2004b. Multi resources GIS hydrologic framework for TMDL assessment. Poster presentation at the 2004 AWRA Spring Specialty Conf.: Geographic Information Systems (GIS) and Water Resources III. Middleburg, Va.: American Water Resources Association.

Di Luzio, M., R. Srinivasan, and J. G. Arnold. 2004. A GIS-coupled hydrological model system for the watershed assessment of agricultural nonpoint and point sources of pollution. *Trans. GIS* 8(1): 113-136.

Doherty, J., and R. J. Hunt. 2009. Two statistics for evaluating parameter identifiability and error reduction. *J. Hydrol.* 366: 119-127.

Doherty, J., and B. E. Skahill. 2006. An advanced regularization methodology for use in watershed model calibration. *J. Hydrol.* 327(3-4): 564-577.

Faber, N. M., and R. Rajko. 2007. How to avoid over-fitting in multivariate calibration: The conventional validation approach and an alternative. *Analytica Chimica Acta* 595(1-2): 98-106.

Gesch, D. B. 2007. The national elevation dataset. In *Digital Elevation Model Technologies and Applications: The DEM Users Manual*, 99-118. 2nd ed. D. Maune, ed. Bethesda, Md.: American Society for Photogrammetry and Remote Sensing.

Gesch, D., M. Oimoen, S. Greenlee, C. Nelson, M. Steuck, and D. Tyler. 2002. The national elevation dataset. *Photogram. Eng. and Remote Sensing* 68(1): 5-11.

Gupta, H. V., S. Sorooshian, and P. O. Yapo. 1999. Status of automatic calibration for hydrologic models: Comparison with multilevel expert calibration. *J. Hydrol. Eng.* 4(2): 135.

Hinnell, A. C., T. P. A. Ferre, J. A. Vrugt, J. A. Huisman, S. Moysey, J. Rings, and M. B. Kowalsky. 2009. Improved extraction of hydrologic information from geophysical data through coupled hydrogeophysical inversion. *Water Resources Res.* 46(11): W00D40, DOI: 10.1029/2009WR008288.

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

Marseguerra, M., E. Zio, and L. Podofillini. 2003. Model parameters estimation and sensitivity by genetic algorithms. *Annals of Nuclear Energy* 30(14): 1437-1456.

Moriasi, D. N., J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith. 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Trans. ASABE* 50(3): 885-900.

Nash, J. E., and J. V. Sutcliffe. 1970. River flow forecasting through conceptual models: Part I. A discussion of principles. *J. Hydrol.* 10(3): 282-290.

Neitsch, S. L., J. G. Arnold, J. R. Kiniry, and J. R. Williams. 2005. *Soil and Water Assessment Tool Theoretical Documentation, Version 2005*. Temple, Tex.: USDA-ARS Grassland, Soil and Water Research Laboratory.

Pokhrel, P., and H. V. Gupta. 2009. On the use of spatial-regularization strategies to improve calibration of distributed watershed models. *Water Resources Res.* 46(1): W01505, DOI: 10.1029/2009WR008066, in press.

Pokhrel, P., H. V. Gupta, and T. Wagener. 2008. A spatial regularization approach to parameter estimation for a distributed watershed model. *Water Resources Res.* 44(12): W12419, DOI: 12410.11029/12007WR006615.

Pokhrel, P., K. K. Yilmaz, and H. V. Gupta. 2009. Multiple-criteria calibration of a distributed watershed model using spatial regularization and response signatures. *J. Hydrol.* DOI: 10.1016/j.jhydrol.2008.12.004 (in press).

R Development Core Team. 2007. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. Available at: www.R-project.org.

Radtke, P. V. W. and T. Wong. 2006. An evaluation of over-fit control strategies for multi-objective evolutionary optimization. In *Proc. IEEE Intl. Joint Conf. on Neural Networks*, 3327-3334. Piscataway, N.J.: IEEE.

Reed, S., and D. Maidment. 1999. Coordinate transformations for using NEXRAD data in GIS-based hydrologic modeling. *J. Hydrol. Eng.* 4(2): 174-182.

Reed, S., V. Koren, M. Smith, Z. Zhang, F. Moreda, D. Seo, and DMIP participants. 2004. Overall distributed model intercomparison project results. *J. Hydrol.* 298: 27-60.

Reiczigel, J., I. Zakariás, and L. Rózsa. 2005. A bootstrap test of stochastic equality of two populations. *American Statistician* 59(2): 1-6.

Schoups, G., N. C. van de Giesen, and H. H. G. Savenije. 2008. Model complexity control for hydrologic prediction. *Water Resources Res.* 44: W00B03, DOI: 10.1029/2008WR006836.

Scott, D. W. 1992. *Multivariate Density Estimation: Theory, Practice, and Visualization.* New York, N.Y.: John Wiley and Sons.

Smith, M. B., D.-J. Seo, V. I. Koren, S. Reed, Z. Zhang, Q.-Y. Duan, F. Moreda, and S. Cong. 2004. The distributed model intercomparison project (DMIP): Motivation and experiment design. *J. Hydrol.* 298: 4-26.

Smith, M., V. Koren, S. Reed, Z. Zhang, D. J. Seo, F. Moreda, and Z. Cui. 2006. The distributed model intercomparison project: Phase 2. Science plan. Silver Spring, Md.: National Weather Service, Hydrology Laboratory. Available at: www.nws.noaa.gov/oh/hrl/dmip/2/docs/dmip_2_plan_march10_06_update.pdf. Accessed 16 September 2010.

Tikhonov, A. N., and V. Y. Arsenin. 1977. *Solutions of Ill-Posed Problems.* New York, N.Y.: V. H. Winston, distributed by Wiley.

Tonkin, M. J., and J. Doherty. 2005. A hybrid regularized inversion methodology for highly parameterized environmental models. *Water Resources Res.* 41(10): W10412, DOI: 10410.11029/12005WR003995.

USDA. 1992. *State Soil Geographic Database (STATSGO) Data Users' Guide*. Publication 1492. Washington, D.C.: USDA-NRCS.

van Werkhoven, K., T. Wagener, P. Reed, and Y. Tang. 2009. Sensitivity-guided reduction of parametric dimensionality for multi-objective calibration of watershed models. *Advances in Water Resources* 32(8): 1154-1169.

Venables, W. N. and B. D. Ripley. 2002. *Modern Applied Statistics with S-Plus*. 4th ed. New York, N.Y.: Springer.

Vogelmann, J. E., S. M. Howard, L. Yang, C. R. Larson, B. K. Wylie, and N. Van Driel. 2001. Completion of the 1990s National Land Cover Dataset for the conterminous United States from Landsat Thematic Mapper data and ancillary data sources. *J. American Soc. for Photogrammetry and Remote Sensing* 67(6): 650-662.

Vrugt, J. A., C. J. F. ter Braak, H. V. Gupta, and B. A. Robinson. 2009. Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling? *Stochastic Environ. Research and Risk Assessment* 23(7): 1011-1026.

Wagener, T., M. Sivapalan, P. A. Troch, and R. Woods. 2007. Catchment classification and hydrologic similarity. *Geogr. Compass* 1(4): 901-931.