Multicriteria parameter estimation for models of stream chemical composition

T. Meixner
Department of Environmental Sciences, University of California, Riverside, California, USA

L. A. Bastidas, H. V. Gupta, and R. C. Bales
Department of Hydrology and Water Resources, University of Arizona, Tucson, Arizona, USA

Received 22 November 2000; revised 1 October 2001; accepted 1 October 2001; published 28 March 2002

[1] The inability to develop an accurate and precise parameter estimation method for catchment hydrochemical models has been a persistent problem. We investigate the use of multicriteria calibration techniques and the selection of the criteria as a first step in solving this problem. We applied a multicriteria search algorithm to the Alpine Hydrochemical Model (AHM) of the Emerald Lake watershed, Sequoia National Park, California. A total of 21 chemical and hydrologic criteria were available for determining model performance. Four subsets of these criteria were selected for the multicriteria analysis using three different methods. The first set used the four least correlated observations of stream chemical composition. A second set of criteria was determined by using the four species with the least correlated root mean square error criteria values (< 0.05 for each pair). Finally, two sets were chosen using the results of a multicriteria sensitivity analysis. The most accurate and precise results were observed using criteria selected using results from a sensitivity analysis, with the correlation analyses being a poor method for selecting criteria. This set of criteria emphasizes attributes of the model structure, the observations, and our understanding of the processes influencing watershed hydrology and water quality. Also, our results gave improved estimates of several hydrologic and biogeochemical processes in addition to identifying a flaw in the current representation of mineral weathering within the AHM, as applied to the Emerald Lake watershed. INDEX TERMS: 1045 Geochemistry: Low-temperature geochemistry; 1863 Hydrology: Snow and ice (1827); 1871 Hydrology: Surface water quality; 1803 Hydrology: Anthropogenic effects; KEYWORDS: multicriteria, model calibration, catchment biogeochemistry

1. Introduction

[2] Over the past several decades, increasingly robust models of stream chemical composition have been developed to address issues ranging from acidification, chemical denudation rates, non-point source pollution, climate change, and land use change [Christophersen et al., 1993; Wolford et al., 1996; Cosby et al., 1985]. These models of catchment hydrology and hydrochemistry are uncoupled from rigorous hypothesis testing because calibration data do not contain enough information to uniquely determine model parameters and there is not a one-to-one relationship between state variables in catchment biogeochemical models and field observations (e.g., the model represents average field conditions, while field measurements are point-scale measurements) [Christophersen et al., 1993]. The decoupling of these models from rigorous hypothesis testing limits their use to comparative studies between models or between catchments with the same model [Bassett, 1997]. Without improved methods of model calibration and validation the useful application of catchment hydrochemical models will remain limited.

[3] In the last several years a number of researchers have focused on the use of multiple fluxes and measurements (including a variety of geochemical tracers and distributed soil moisture) for automatic model calibration and validation of hydrochemical models. In one of the first examples of the multicriteria analysis of a hydrochemical model, Hooper et al. [1988] and de Grosbois et al. [1988] illustrated some deficiencies in the Birkenes model using contrasting simulations of $\delta^{18}$O and discharge. Later applications showed that while these model structural flaws could be overcome, good simulation of hydrochemical fluxes would be sacrificed [Stone and Seip, 1989; Lundquist et al., 1990]. In a more recent application, Kuczera and Mroczkowski [1998] investigated the use of monthly streamflow, stream chloride, and groundwater levels to determine which of two model structures was best for modeling a catchment. They determined that groundwater stage was not useful, while stream chloride concentration was useful in determining the right structure of the model. There are three lessons from these past studies. First, hydrochemical data can be used to calibrate the hydrologic parameters within a model. Second, because of model structure errors and data errors, optimal solutions for chemistry and other fluxes may conflict. Third, these methods can uncover flaws in model structure and point us in the direction of how to improve the model.

[4] This paper builds upon recent developments in the multicriteria analysis of hydrologic models [Gupta et al., 1998; Bastidas et al., 1999; Gupta et al., 1999]. Multicriteria methods provide a tool that can be used to solve the calibration, parameter estimation, sensitivity, and model evaluation problems in catchment hydrochemical modeling. In this manuscript we discuss how to select the criteria for multicriteria calibration of hydrochemical models and some of the implications of these results for hydrochemical modeling.

[5] We addressed two main questions by conducting a multicriteria calibration of the Alpine Hydrochemical Model (AHM) of...
2. Methods

2.1. Multicriteria Parameter Estimation Methodology

A thorough discussion of the application of multicriteria theory to calibration of conceptual physically based models is given by Gupta et al. [1998]. The following is a brief summary of that methodology. Consider a model with parameters \( \Theta = \{ \Theta_1, \ldots, \Theta_q \} \) that is to be calibrated with observations \( O_j \) over \( m \) simulated model output variables. For each simulated response \( X_j \) it is possible to define a criterion \( f_j(\Theta) \) that represent the distance between the simulated value \( X_j \) and the observations \( O_j \). The criterion \( f_j \) may be defined with any number of measures of model error or residual. The root mean square error (RMSE) is commonly used since its units are the same as those of the observations. RMSE can be represented as

\[
RMSE(q) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (O_i - X_i(\theta))^2},
\]

where \( q \) is discharge, \( \theta \) is the set of model parameters, \( n \) is the total number of observations \( O \), and \( X \) is the simulated value at time step \( t \). The multicriteria model calibration problem can formally be stated as

\[
\min F(\theta) = \{ f_1(\theta), \ldots, f_m(\theta) \} \quad \theta \subset \Theta,
\]

where the goal is to find values for \( \theta \) within the feasible set \( \Theta \) that minimize all of the criteria \( (f_j(\Theta), \quad l = 1, \ldots, m) \) simultaneously.

In practice, it is not possible to minimize all criteria simultaneously. Instead, a set of solutions is commonly found, with this set of solutions having the property that it is necessary to deteriorate performance of one criterion in order to improve the performance of a second criterion within the set of solutions. Figure 1 illustrates a multicriteria problem with two parameters \( (\theta_1, \theta_2) \) and three criteria \( (f_1, f_2, \text{and } f_3) \). The Pareto set represents the set of solutions that can objectively be considered to be better than all other possible solutions. However, comparisons within the set of solutions are not possible on an objective basis, and therefore all of the solutions must be compared as a whole to other possible Pareto set solutions for a given model. The Pareto set is sometimes called the trade-off set, noninferior set, nondominated set, or the efficient set. The Pareto set represents the set of solutions that can objectively be considered to be better than all other possible solutions. However, comparisons within the set of solutions are not possible on an objective basis, and therefore all of the solutions must be compared as a whole to other possible Pareto set solutions for a given model. The Pareto set is related to errors in model structure and the calibration data set. Only when a perfect model and perfect data are available will the Pareto set be a unique solution.

Yapo [1996] presented an efficient population-based optimization strategy that provides an approximate representation of the Pareto set with a single optimization run. This algorithm, Multiobjective Complex Evolution (MOCOM-University of Arizona (UA)), is based on the Shuffled Complex Evolution (SCE-UA) optimization method [Duan et al., 1992, 1993]. The MOCOM-UA method begins by sampling the feasible space \( \Theta \)}
at a number of locations. At each location the multicriteria vector $F(\Theta)$ is computed, and the population is ranked and sorted using Pareto rank [Goldberg, 1989]. Simplexes of $n + 1$ points are then selected using a rank-based method. Each simplex is evolved in an improvement direction using a multicriteria extension of the downhill simplex method. Iterative application of the ranking and simplex evolution steps causes the population to converge toward the Pareto set. The algorithm terminates automatically when all points in the sample become mutually nondominated [Yapo et al., 1998].

2.2. Site Description

[9] The current evaluation of AHM builds on an application of the model to the 1985–1987 intensive field campaign at ELW [Tonnesen, 1991; Wolford et al., 1996]. ELW is a 120 ha headwater catchment located in the Sierra Nevada (36°35′N, 118°40′W), with elevation ranging from 2800 m at the lake to 3417 m at the summit of Alta Peak. The watershed is 48% covered by exposed granite and granodiorite, 23% covered by soil, and 23% covered by talus and includes a 2 ha lake. On average, snowfall represents 95% of total annual precipitation. Streamflow is dilute, with conductivity ranging from 2 to 10 µS cm$^{-1}$ and alkalinity ranging from 15 to 50 µequiv L$^{-1}$. The watershed is considered to be sensitive to changes in climate and atmospheric deposition because of thin soils, dilute waters, and snow-dominated hydrology. The 1986 water year (October 1985 to September 1986) had significantly above-normal snowfall, and 1987 had significantly below-normal snowfall. A simulation over the 2 years thus represents a robust test of the model since it must perform well in both wet and dry periods.

2.3. Model

[10] AHM is a lumped conceptual model for alpine watersheds that requires the specification of the hydrologic and biogeochemical processes occurring in a watershed [Wolford et al., 1996]. The ELW was partitioned into three terrestrial subunits (rock, talus, and soil), a stream, and a lake (Figure 2). Each terrestrial subunit is made up of different compartments representing the snowpack, snowpack-free water, snowmelt, surface runoff, and interception by trees and litter and may contain multiple soil horizons. Stream and lake subunits have compartments for the snowpack, snowpack-free water, snowmelt, stream/lake ice, and (as appropriate) streamflow or a stratified lake. Within this structure a set of parameters define the routing of flow from the rock subunit to the talus and soil subunits and from there sequentially into the stream and the lake and out of the watershed. AHM adjusts hydrologic and chemical inputs, outputs, and state variables for 13 separate compartments representing snow, vegetation, infiltration, and soil processes. AHM calculates chemical equilibrium and moves water and chemicals between compartments on a daily time step, with some processes calculated on a subdaily time step. Model output can include detailed descriptions of all chemical calculations, tracking of both chemical and hydrologic storage and changes in storage within the watershed, soil chemical concentrations, and stream concentrations.

[11] Meixner et al. [1999] conducted a multicriteria sensitivity analysis to determine how model parameters impacted the various criteria of model performance for the ELW. The results of the sensitivity analysis also provided information about the dominant processes in the watershed. For the sensitivity analysis, 20,000 Monte Carlo simulations were conducted during which all 24 model parameters were varied independently of each other (Table 1). A total of 21 criteria were evaluated using the RMSE function during each of these simulations. The criteria were as follows: 1 criterion for discharge, 10 criteria for the difference between modeled and observed stream concentrations for 10 chemical species ($Ca^{2+}$, $Mg^{2+}$, $Na^+$, $K^+$, $Cl^-$, $SO_4^{2-}$, $NO_3^-$, acid neutralizing capacity (ANC), pH, and Si), and 10 criteria based on the difference between modeled and observed daily mass flux for the same chemical species. Daily mass flux was calculated as the product of discharge and chemical concentration.

[12] Next, the results were analyzed to determine parameter sensitivity using the multiobjective generalized sensitivity analysis algorithm (MOGSA) [Bastidas et al., 1999]. This algorithm uses bootstrapping and Pareto rank techniques to determine parameter sensitivity with either a single objective or multiple objectives. MOGSA was used to determine parameter sensitivity for all 24 parameters for the 21 criteria individually and for a global set of criteria. A key strength of the sensitivity analysis is the ability to identify which criteria are affected by which parameters; by extension the sensitivity results also do the reverse: They indicate which criteria provide information about model parameters and watershed processes.

[13] The sensitivity results indicated that mass flux and concentration criteria identified different physical processes in the watershed. For example, chemical concentration criteria showed more sensitivity to mineral weathering parameters, while mass flux criteria showed more sensitivity to cation exchange processes. The results also indicated the importance of the accurate character-
relationship between used in the sensitivity analysis. In the models current form A and B magnitude and shape of the ionic pulse, and initial concentration in the snowpack, coefficients A, B, and D define the density, and specific surface area for soil particles. The weathering coefficients varied independently for each subunit within the same mathematical range. The remaining two sets of criteria were selected on the basis of information gained from the sensitivity analysis. The two sets were chosen based on which criteria shared the least number of sensitive model parameters. These two sets consisted of Q, Ca\(^{2+}\), SO\(_4^{2-}\), and Cl\(^{-}\) criteria. In one case the concentration criteria of SO\(_4^{2-}\) and Cl\(^{-}\) were used; in the other the mass flux criteria were used in both of these sets. The case using only concentration criteria will be referred to as the sensitivity-concentration case. The case using a mix of concentration and mass flux criteria as determined by the sensitivity analysis will be referred to as the sensitivity-mass flux case. The equation used to represent snowpack elution in the Alpine Hydrochemical Model is \( C/C_{ave} = A \times B \times e^{-\beta} \times \sum_{i=1}^{N} \left( 1 - \alpha \right) \times D 	imes Z^{-D} \times x^{-D} \), where \( C \) is the solute concentration in snowmelt, \( C_{ave} \) is the initial concentration in the snowpack, coefficients A, B, and D define the magnitude and shape of the ionic pulse, and X is the fraction of the snowpack that has already melted. In the models current form A and B are small, so D is the dominant parameter and was thus used in the hydraulic conductivity: \( K_s = K_{sat} \sqrt{W \left( 1 - (1 - W)^{1/\alpha} \right)^{2/3}} \), where \( K_s \) is the unsaturated hydraulic conductivity, \( K_{sat} \) is the saturated hydraulic conductivity, \( W \) is (\( 0 - \theta_n \))/(\( \theta_{sat} - \theta_n \)), \( n \) is a parameter that defines the relationship between \( 0 \) and \( \theta_n \), the shape of the curve as water content declines, \( \theta_n \) is the water content of the soil, \( \theta_s \) is the water content at which no flow will occur, and \( \theta_{sat} \) is the saturated water-holding capacity of the soil. The equation for representative of the Emerald outflow. The correlation analysis showed that the four least correlated time series were discharge (\( Q \)) and the concentrations of H\(^+\), Mg\(^{2+}\), and Si (\( p < 0.6 \) for each pair). This set of criteria will be referred to as the data correlation case. A second set of four criteria was determined by using the four species with the least correlated RMSE values (\( p < 0.05 \) for each pair). The RMSE values were calculated as part of the sensitivity analysis [Meixner et al., 1999]. These four criteria were \( Q \), ANC concentration, SO\(_4^{2-}\) mass, and Cl\(^{-}\) mass. This set will be referred to as the criteria correlation case. The log K is for adsorption of SO\(_4^{2-}\) and H\(_2\)SiO\(_4\). Exchange reactions for each were as follows: SO\(_4^{2-}\) + Y + 2H\(^+\) \rightleftharpoons YH\(_2\)SO\(_4\) and H\(_2\)SiO\(_4\) + Z \rightleftharpoons Z - H\(_2\)SiO\(_4\). Total site concentrations from Emerald Lake watershed optimization are used here. These parameters cannot be varied independently for the soil and talus subunit. While modeled as an adsorption reaction, H\(_2\)SiO\(_4\) adsorption is functionally equivalent to the precipitation of amorphous Si. The partial pressure of carbon dioxide pCO\(_2\) was varied simultaneously for both subunits. The weathering coefficients \( \alpha \) and \( \alpha \) were varied independently of each other but uniformly for all species because of issues of charge balance. They contribute to weathering via the equation mol = \( 4 \times n \times [H^+] \), where mol is moles added to the subunit, A is total area of the surfaces involved in reactions. [H\(^+\)] is hydrogen ion concentration, and \( \alpha \) and \( \alpha \) are constants. The total surface area is determined from available field data (Table 1) [Meixner et al., 1999]. The AHM model of the ELW simulates a total of 11 fluxes that can be readily compared with available stream observations. Since criteria containing similar information should not be used simultaneously in a multicriteria optimization problem [Gupta et al., 1998], we developed four sets of complementary criteria; each set contained four criteria; using more than four criteria or fewer than four criteria gave inferior results [Meixner, 1999]. The first set was chosen on the basis of a correlation analysis of the observations of stream chemical composition at the Emerald outflow. The correlation analysis showed that the four least correlated time series were discharge (\( Q \)) and the concentrations of H\(^+\), Mg\(^{2+}\), and Si (\( p < 0.6 \) for each pair). This set of criteria will be referred to as the data correlation case. A second set of four criteria was determined by using the four species with the least correlated RMSE values (\( p < 0.05 \) for each pair). The RMSE values were calculated as part of the sensitivity analysis [Meixner et al., 1999]. These four criteria were \( Q \), ANC concentration, SO\(_4^{2-}\) mass, and Cl\(^{-}\) mass. This set will be referred to as the criteria correlation case. For each of these sets of criteria the Pareto set was estimated using the MOCOM-UA algorithm. We selected a population size of 250 for the multicriteria search [Meixner, 1999]; each calibration case thus has 250 different final parameter sets that are considered acceptable. Calibration success was compared using the Nash-Sutcliffe values [Nash and Sutcliffe, 1970] for the criteria used and not used during the various calibrations. The Nash-Sutcliffe statistic is

\[
E = 1 - \frac{\sum_{i=1}^{N} (O_i - P_i)^2}{\sum_{i=1}^{N} (O_i - \bar{O})^2},
\]

where \( O \) is the observation, \( \bar{O} \) is the mean observation, \( P \) is the predicted value, \( i \) is the observation or prediction in question, and \( N \) is the total number of observations. A Nash-Sutcliffe value greater than zero indicates that the model is superior to the observational mean as a predictor of stream chemical composition [Legates and McCabe, 1999]. The calibration results for each of the four sets of criteria used in the MOCOM-UA algorithm will be viewed in four different ways: parameter space, criteria space, comparisons of the model results to the mean (using the Nash-Sutcliffe statistic), and a comparison of the results to those of Wolford et al. [1996] (referred to as the default parameter results). The set of criteria that performs best is the one that has the smallest variation in estimated parameters and gives a calibrated model that is superior to standard predictors of stream chemical composition.

### 3. Results

[18] The correlation case, the sensitivity-concentration case, and the sensitivity–mass flux case resulted in relatively small parameter spaces for the final 250 Pareto set results. The parameter values from the data correlation case spanned most of the possible...
have some interesting contrasts. The case using only concentration would result from using 250 sets of random parameter values. There is little coherence of the results, and they are similar to what criteria was significantly worse for Mg\textsuperscript{2+} either as mass flux or correlation case the other calibration sets result in lower values for data correlation case were inferior to those of the other three cases.

Comparing the overall results for each of the methodologies to the mean indicates, first, that on the basis of several criteria the current performance of the model is inadequate at best (Table 2). In particular, the mean is a superior predictor for Cl\textsuperscript{-}, Si, Na\textsuperscript{+}, and NO\textsubscript{3}\textsuperscript{-} for all but a few of the 1000 “best” simulations (250 parameter sets for each set of criteria) between the four sets of criteria used here. In general, the model does a very good job of simulating the mass flux of all species, although it does not simulate the mass flux of Cl\textsuperscript{-} as well as it simulates the other mass flux criteria. The model is expected to perform much better than the mean mass flux since mass flux varies by several orders of magnitude because of wide variations in flow, while observed concentrations vary by a maximum of 1 order of magnitude. Among the four sets the sensitivity-concentration case does the best on a mean basis with the average model prediction superior to the observational mean 178 out of 250 times. The worst of the four is the data correlation case, with the model being better than the mean as a predictor of stream chemical composition only 150 out of 250 times.

Comparing the results to the default parameters indicates that the best of the four criteria selection methodologies is the sensitivity-concentration case (162 out of 250 simulations improved the simulations for several criteria, with Mg\textsuperscript{2+}, Na\textsuperscript{+}, NO\textsubscript{3}\textsuperscript{-}, and H\textsuperscript{+} predictions not improved in any significant way over the default parameter results. The results for the sensitivity–mass flux case were not nearly as good, with several criteria not being improved with respect to the default parameter results. These results indicate that choosing criteria using the correlation or lack thereof of observations is not a good methodology for selecting the criteria used for multicriteria calibration since results for the data correlation case were inferior to those of the other three cases.

The criteria space results for the 250 best parameter sets for the four different criteria selection cases provide additional information on how each of the cases performed (Figure 5). The criteria correlation case results show improvement over the results of the default parameters for ANC, Na\textsuperscript{+}, Cl\textsuperscript{-}, and Si, while this set of criteria was significantly worse for Mg\textsuperscript{2+} either as mass flux or concentration. The results for the data correlation case show that for each of the criteria some of the 250 simulations are better than those of the default parameters, while others are much worse. There is little coherence of the results, and they are similar to what would result from using 250 sets of random parameter values. The results for both sensitivity analysis cases have some interesting contrasts. The case using only concentration

Figure 3. Hydrologic parameter ranges for four different calibration cases. See Table 1 for a description of the parameters. Soil and talus labels refer to soil and talus subunits in AHM model of Emerald Lake (Figure 2). Each plot of Figure 3 is for a single-criteria case that we investigated. The bold line represents the default parameter values [Wolford et al., 1996]. Shaded lines represent the 250 Pareto solutions for each calibration case. The y axis value is a multiple of the default parameter value [Wolford et al., 1996]. The evapotranspiration (ET) parameter for both subunits for all calibration cases has a lower value than the default parameter value [Wolford et al., 1996]. Except for the data correlation case the other calibration sets result in lower values for the parameter K\textsubscript{sat} in the soil subunit. A larger volume of talus than currently parameterized is indicated by higher values for talus depth and theta parameters.

Figure 4. Chemical parameter ranges for four different calibration cases. Description of Figure 4 is the same as for Figure 3 except ranges for the 14 chemical parameters are shown. The plots all indicate that the elution parameter (EL) should be lower than currently parameterized. There is a conflict in the calibration results with the criteria correlation case having a lower value for the weathering parameter \( \alpha \), while both of the sensitivity analysis cases have a higher value for \( \alpha \).
improved) (Table 3). There were several criteria that none of the methodologies were able to improve, including H+ and NO3- mass flux. There are several other items of interest apparent in the results. First, neither of the sensitivity analysis cases was able to result in a significantly improved simulation of Na+ and NO3- concentration or mass flux. Second, neither of the correlation-based criteria sets was able to improve the simulation of SO4\(^{2-}\) for either concentration or mass flux. Third, for the criteria correlation case, almost all 250 simulations had inferior predictions for Ca\(^{2+}\) and Mg\(^{2+}\) concentration or mass flux. Finally, the two correlation-based criteria selection methodologies were significantly inferior to the two cases where criteria were selected using the results of a sensitivity analysis.

4. Discussion

4.1. Which Set of Criteria Is Best?

[22] According to the current analysis the four criteria that are best suited to this first step of multicriteria analysis are Q and the concentration of Ca, SO4\(^{2-}\), and Cl\(^{-}\). The companion set of criteria using the mass flux of SO4\(^{2-}\) and Cl\(^{-}\) was inferior because it did not improve model results for all criteria as much as the concentration case did. Both the data correlation case and criteria correlation case were inferior to the sensitivity-concentration case. The criteria correlation results had inferior simulations of the 21 criteria modeled using either the mean or the default parameters as the benchmark for model comparison. The data correlation case resulted in little improvement in model performance over the two available benchmarks and had a parameter space that spanned the feasible space and thus contained little to no information about model or watershed processes.

4.2. How to Choose Criteria?

[23] Criteria selected using multicriteria sensitivity analysis resulted in the best calibrated models, suggesting selection should consider which criteria give the most independent parameter sensitivity results. The difference between the sensitivity-mass flux case and the concentration-concentration case illustrates that criteria selection should be limited to measured properties as opposed to transformation of the observed data time series, confirming previous results by others [Bastidas et al., 1999; Gupta et al., 1999] but extending that conclusion to water quality models.

[24] Selecting criteria using sensitivity analysis results incorporates information contained in the data and the model and the user’s intuition. The sensitivity analysis incorporates information from the data since the data are used to determine which simulations are considered behavioral and which are considered non-behavioral. Information from the model is obviously incorporated since the purpose of the sensitivity analysis is to determine which criteria are sensitive to which model parameters [Bastidas et al., 1999]. The user’s intuition is involved since the user must determine which criteria contain information about the model form and the processes represented by the model. The user does this by using the sensitivity results and his/her practical understanding of the important processes to determine which set of criteria would be best.

[25] Further improvements in criteria selection should be sought, since the current method of selecting criteria is still rather subjective. Additional statistical analysis of criteria and parameter values, resulting from Monte Carlo simulations, may reveal more objective ways to select criteria. These analysis tools might include regression tree analysis, principal component analysis, and cluster analysis of model criteria and parameter values. Preliminary results with a cluster analysis of the criteria values from 20,000 Monte Carlo simulations [Meixner et al., 1999] supported the four criteria selected using the sensitivity analysis. However, the cluster analysis did not provide a single unique result. Using other sets of criteria that would be acceptable according to the cluster analysis did not result in superior model simulations. This failure was due to a conflict in model structure related to mineral weathering which will be discussed in more depth. This failure of cluster analysis indicates that more work is needed to determine the best methodology in selecting criteria for use in the multicriteria analysis of hydrologic and hydrochemical models.

4.3. What Did We Learn About the Hydrologic and Chemical Processes?

[26] The results reveal several things about the processes controlling stream chemical composition in the Emerald Lake watershed. First, the results indicate that evapotranspiration is less than currently modeled (Figure 3) using default parameters. Field and modeling exercises should focus on summertime estimates of ET, since model predictions of stream chemical composition are insensitive to ET parameters during snowmelt [Meixner et al., 1999]. Second, the results indicate that the hydraulic conductivity (Ksat) of the soil should be higher than the current value; for the talus, results were mixed (Figure 3). Also, the results point to a deeper talus and a larger saturated water-holding capacity (θ) than in the default parameters. These results indicate the need for a larger reactive volume for the talus subunit than currently used in the model.

[27] For the chemical parameters, there are fewer clear-cut examples of the analyses ability to limit the parameter space and
to understand the processes controlling stream chemical composition. One exception is that all cases except the data correlation case indicated a lower elution parameter (EL). Additionally, it should be noted that the model generally performed poorly in simulating Cl− concentrations. This is surprising considering that Cl− is a conservative ion with only hydrologic parameters and the ion elution parameter has an influence on the model’s ability to simulate Cl−. The reason for the poor performance in simulating Cl− concentration may be that the real ionic pulse varies from year to year according to snowpack conditions [Harrington and Bales, 1998a, 1998b]. In the current model application the ionic pulse was held constant for both years.

[28] Finally, the two correlation cases indicated a slower weathering rate, while the two sensitivity analysis cases indicated a higher rate. These results might be caused by parameter interaction between k and α. However, there is no discernable relationship for all four cases between the two parameters (ρ < 0.02). The mixed results may arise because the criteria behave differently. Silica, Na+, and ANC would be better simulated by a lower value for the weathering rate parameter α, while cations Ca2+, Mg2+, and Na+ need a higher α.

4.4. Model Structure

[29] This conflict represents a problem within the current model structure for mineral weathering. Three criteria, ANC, Si, and Na+, had mean, median, and at times maximum Nash-Sutcliffe values (out of the 250 Pareto results) less than zero for most of the cases; thus, for these three criteria the mean is a better predictor of stream chemical composition than is the model. This result, together with the conflict between modeling ANC, Si, and Na+ versus Ca2+, Mg2+, and K+, points to a flaw in model structure. The best simulations of Si occurred when ANC was the only weathering product included in the set of four criteria. When Si was one of the criteria, the results were not nearly as good, possibly because of a conflict in parameter estimates since Mg2+ was also one of the geochemical criteria, e.g., ANC or Ca2+) since it has been indicated that all weathering product criteria and either ANC, Na+, and Si criteria or Ca2+, Mg2+, and K+ criteria. For example, when calibrating on Q, Cl−, Ca2+, and K+, results were poor for ANC, Na+, and Si.

[30] Three other sets of criteria simulated based on a cluster analysis of criteria values gave similar results. For these other calibration cases the only good results were for those cases that included only nonweathering product criteria and either ANC, Na+, and Si criteria or Ca2+, Mg2+, and K+ criteria. For example, when calibrating on Q, Cl−, Ca2+, and K+, results were poor for ANC, Na+, and Si.

[31] These results indicate a mistake in model structure either in the temporal variability or stoichiometry of mineral weathering. Currently, weathering is assumed to occur evenly throughout the year, an assumption confirmed by Shaw [1997]. However, a different representation of weathering, such as equilibrium dissolution of kaolinite into solution as suggested by Campbell et al. [1995] might be a better approach than the current simple kinetic approach adopted by Wolford [1992].

4.5. Broader Implications to Multicriteria Analysis

[32] Our results illustrate the sort of problems that need to be addressed to successfully apply multicriteria analysis. First, our method shows that criteria not used in the calibration can be used to determine the success of the calibration procedure, and in our case this methodology was successful in determining which set of criteria was best suited for use in a multicriteria analysis. Second, our results illustrate that in addition to the classic problem of parameter interaction and parameter correlation, there is now the problem of correlation of criteria and criteria interaction. This problem may explain some of the difficulties others have had in using multicriteria analysis. For example, de Grosbois et al. [1988] and Hooper et al. [1988] had to use several different weighting schemes to find the right one to permit better parameter identification with a simple two-criteria case. Some of the problems they had in their analysis of the Birkenes model may have been solved if they had tools such as MOCOM-UA available and they had investigated the interaction of criteria more fully (incorporating several geochemical criteria, e.g., ANC or Ca2+) since it has been shown that stream chemical composition contains information about watershed hydrology [Hooper et al., 1990]. The use of this additional information would have helped determine how model
structure could be improved. Such an analysis may have prevented the later conflicts between the improvements they made to model hydrologic structure and the deterioration in simulating stream chemical composition when their results were extended by others [Stone and Seip, 1989; Lundquist et al., 1990]. Our results also indicate that good scientific intuition is the reason behind the past success with multicriteria model analysis [de Grosois et al., 1988; Hooper et al., 1988; Uhlenbrook and Leibundgut, 2000]. These previous investigations all used external information from the literature or from their personal experience with the data set and model at hand to select the multiple criteria they used for calibration.

Looking at the results of Mroczkowski et al. [1997] and Kuczera and Mroczkowski [1998] in light of ours, we might infer that groundwater stage and stream discharge series contain conflicting information about watershed processes as represented by their model and thus the lack of information that combining discharge and groundwater stage in a single calibration run provided in their analysis. As did Kuczera and Mroczkowski [1998], we found that several different criteria combinations in our analysis did not improve model performance in any significant way. Their warning should be reiterated: There is a danger in “assuming that augmenting streamflow data with other response time series data will significantly reduce parameter uncertainty.” This warning should be expanded to include hydrochemical models and biogeochemical models. Because of the cost of acquiring time series other than streamflow, they recommend better studies a priori of the worth of additional time series of data. We would add to this caution that different combinations of criteria may improve results in different and possibly conflicting ways. These combinations should be used to investigate model structure and the relative worth of data time series.

Our results also indicate that when doing manual calibration using visual or numerical comparison of model results to adjust parameters, one should be aware of and use the multicriteria nature of the calibration problem as it pertains to hydrochemical catchment models. Undoubtedly, individuals who manually calibrate water quality models of all varieties already include knowledge of the multicriteria nature of their calibration problem through human interaction with the model.

In addition, we can use our results to confirm or deny the validity of existing manual calibration methodologies. For example, in an elaborate manual calibration of the Model of Acidification of Groundwater in Catchments (MAGIC) [Cosby et al., 1985; Church and Van Sickle [1999] determined 10 separate parameter sets for 36 catchments in the Adirondack Mountains, New York. After determining the 10 parameter sets, they used the median for each parameter to predict the future chemistry of each of the 36 catchments. Is it an acceptable practice to use the median parameter value from a set of 10 or in our case 250 different calibration results? To answer this question, we compared simulation results using the median parameter values of the 250 parameter sets for the four cases we studied to the 250 criteria space results for each case. The results by case and criteria are summarized in Table 4, which shows the fraction (0.0–1.0) of Pareto results that performed worse than the median for each of the 21 criteria and each of the four cases we investigated. A value of 0 indicates that the median parameter values performed worse than every member of the Pareto set. A value of 1.0 means that the median parameter values performed better than all members of the Pareto set. On average, the median parameter values performed better than the Pareto solutions 43% of the time. This result indicates that the median of the parameter values for each case typically performed a little worse than the median of the Pareto results. The median does not perform equally well for all criteria. For example, the median is significantly inferior to the Pareto results for NO$_3^-$ but is significantly superior to the Pareto results for Si.

The two cases that had the best performance for the median parameter values were the data correlation case and the sensitivity analysis–concentration case. The result for the data correlation case is expected since the parameter and criteria values used for that case were similar to the results for 250 parameter sets generated with random numbers. Any parameter set might be expected to perform well in these circumstances. However, the sensitivity analysis–concentration case results indicate that for a properly conducted multicriteria calibration (either automatic or manual) the median represents a reasonable way to arrive at a preferred solution from the Pareto set results. Therefore, in circumstances where a competent multicriteria calibration has been conducted, using the median is an acceptable means of determining a single preferred solution to be used for simulations of perturbed conditions.

<table>
<thead>
<tr>
<th>Method</th>
<th>Q</th>
<th>H$^+$</th>
<th>ANC</th>
<th>Ca$^{2+}$</th>
<th>Mg$^{2+}$</th>
<th>K$^+$</th>
<th>Na$^+$</th>
<th>SO$_4^{2-}$</th>
<th>Cl$^-$</th>
<th>Si</th>
<th>NO$_3^-$</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Criteria</td>
<td>0.57</td>
<td>0.36</td>
<td>0.16</td>
<td>0.00</td>
<td>0.21</td>
<td>0.60</td>
<td>0.91</td>
<td>0.74</td>
<td>0.00</td>
<td>0.22</td>
<td>0.00</td>
<td>0.35</td>
</tr>
<tr>
<td>Data</td>
<td>0.78</td>
<td>0.39</td>
<td>0.35</td>
<td>0.17</td>
<td>0.73</td>
<td>0.83</td>
<td>0.47</td>
<td>0.18</td>
<td>1.00</td>
<td>0.00</td>
<td>0.24</td>
<td>0.52</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.44</td>
<td>0.01</td>
<td>0.07</td>
<td>0.20</td>
<td>0.17</td>
<td>1.00</td>
<td>0.80</td>
<td>0.83</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>0.46</td>
</tr>
<tr>
<td>Concentration</td>
<td>0.44</td>
<td>0.03</td>
<td>0.14</td>
<td>0.11</td>
<td>0.31</td>
<td>0.98</td>
<td>0.89</td>
<td>0.96</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>0.40</td>
</tr>
<tr>
<td>Mean</td>
<td>0.56</td>
<td>0.20</td>
<td>0.22</td>
<td>0.16</td>
<td>0.21</td>
<td>0.83</td>
<td>0.86</td>
<td>0.75</td>
<td>0.05</td>
<td>0.81</td>
<td>0.00</td>
<td>0.43</td>
</tr>
</tbody>
</table>

### 5. Conclusions

Four main conclusions can be drawn from our results. First, the model being used, the data, and the user’s intuition (knowledge of each criterion’s connection to processes represented by the model) must be combined to determine what set of criteria to use in a multicriteria analysis. This result has broad implications for the multicriteria analysis and calibration of hydrochemical models. In essence, this result indicates that we need to fully account for our current hypotheses (models) of watershed hydrochemical processes in order to develop hypothesis tests on the correctness of the assumptions built into those models. Scientific intuition should be used to develop hypothesis tests that can use observed data to determine whether our existing hypotheses are correct or incorrect.

Second, fluxes not used in calibration are useful for evaluating calibration results. Noncalibrated fluxes represent an

### Table 4. Fraction of Pareto Results Performing Worse Than Media Parameter Values

<table>
<thead>
<tr>
<th>Method</th>
<th>Q</th>
<th>H$^+$</th>
<th>ANC</th>
<th>Ca$^{2+}$</th>
<th>Mg$^{2+}$</th>
<th>K$^+$</th>
<th>Na$^+$</th>
<th>SO$_4^{2-}$</th>
<th>Cl$^-$</th>
<th>Si</th>
<th>NO$_3^-$</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Criteria</td>
<td>0.57</td>
<td>0.36</td>
<td>0.16</td>
<td>0.00</td>
<td>0.21</td>
<td>0.60</td>
<td>0.91</td>
<td>0.74</td>
<td>0.00</td>
<td>0.22</td>
<td>0.00</td>
<td>0.35</td>
</tr>
<tr>
<td>Data</td>
<td>0.78</td>
<td>0.39</td>
<td>0.35</td>
<td>0.17</td>
<td>0.73</td>
<td>0.83</td>
<td>0.47</td>
<td>0.18</td>
<td>1.00</td>
<td>0.00</td>
<td>0.24</td>
<td>0.52</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.44</td>
<td>0.01</td>
<td>0.07</td>
<td>0.20</td>
<td>0.17</td>
<td>1.00</td>
<td>0.80</td>
<td>0.83</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>0.46</td>
</tr>
<tr>
<td>Concentration</td>
<td>0.44</td>
<td>0.03</td>
<td>0.14</td>
<td>0.11</td>
<td>0.31</td>
<td>0.98</td>
<td>0.89</td>
<td>0.96</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>0.40</td>
</tr>
<tr>
<td>Mean</td>
<td>0.56</td>
<td>0.20</td>
<td>0.22</td>
<td>0.16</td>
<td>0.21</td>
<td>0.83</td>
<td>0.86</td>
<td>0.75</td>
<td>0.05</td>
<td>0.81</td>
<td>0.00</td>
<td>0.43</td>
</tr>
</tbody>
</table>
alternative to the traditional split sample method of evaluating model calibration results and methodologies.

[39] Third, multicriteria parameter estimation is a promising avenue for investigating the accuracy and structure of hydrochemical watershed models. Our multicriteria calibration uncovered a flaw in the representation of mineral weathering. Changes in the way mineral weathering is represented can be tested to see if the alternative representations eliminate the conflict in modeling ANC, Si, and Na⁺ and the rest of the cations. The ability of multicriteria methods to determine flaws in model structure, improve parameter estimates, and discern good and bad models shows much promise, but caution should be used in proceeding as not all criteria contain information that can be used to improve hydrochemical models and parameter estimates. Still, multicriteria techniques are hypothesis-testing tools that promise to increase the utility of catchment hydrochemical models for determining the relative importance of processes internal to watersheds.

[40] Finally, our results have implications beyond model calibration investigations of catchment hydrochemistry. While many different methods of looking at compromise solutions exist and remain to be investigated, the median value for each parameter is a good compromise solution when a single-parameter set is needed for extrapolative solutions. This result confirms an existing simple practice for using a single compromise solution to make extrapolative model simulations.

[41] Acknowledgments. A Canon National Park Science Scholarship for the primary author made this work possible. This work was conducted, in part, for completion of a Ph.D. by the lead author at the University of Arizona. NASA-EOS, NSF, and NOAA research grants provided additional support for this work. Special thanks go to J. Sickman, A. Leydecker, and J. Melack for their help in interpreting field data and to the Surface Water Calibration Research Group at the University of Arizona for the many good conversations and quality insight into model calibration and parameter estimation. A debt of gratitude is owed to S. Sorooshian, whose “Analysis of Hydrologic Systems” course led the primary author to the topic of this paper. M. J. Lees, R. P. Hooper, and an anonymous reviewer are thanked for their comments, which improved this manuscript.

References