Toward improved identification of hydrological models: A diagnostic evaluation of the “*abcd*” monthly water balance model for the conterminous United States

Guillermo F. Martinez[1] and Hoshin V. Gupta[1]

Received 15 June 2009; revised 9 February 2010; accepted 25 February 2010; published 4 August 2010.

[1] Continental-scale water balance (WB) assessments are important for characterizing hydrologic systems and understanding regional-scale dynamics and for identifying hydroclimatic trends and systematic data biases. However, it is not clear whether existing models can reproduce the catchment dynamics observed in nature. Nor has our ability to evaluate model results kept pace with computational and data processing abilities. Consequently, methods for diagnostic model evaluation and improvement remain weak. There is a need for well-conceived, systematic strategies to guide model selection, establish data requirements, estimate parameters, and evaluate and track model performance. We examine these challenges in the context of monthly WB modeling for the conterminous United States by applying the “*abcd*” model to 764 catchments selected for their comprehensive coverage of hydrogeological conditions. By examining diagnostically relevant components of model error, we evaluate the details of its spatial variability across the continental United States. Model performance, parameters, and structures are found to be correlated with hydroclimatic variables. However, our results indicate a need for the conventional identification approach to be improved. Because they do not constrain models to reproduce important hydrological behaviors, reported values of NSE or r^2 performance can be misleading. Further, we must establish suitable model hypotheses with appropriate spatiotemporal scale for each hydroclimatic region. Until these issues are resolved, such models cannot reliably be used to infer the spatiotemporal dynamics of continental-scale water balance or to regionalize model structures and parameters to ungaged locations.

Citation: Martinez, G. F., and H. V. Gupta (2010), Toward improved identification of hydrological models: A diagnostic evaluation of the “*abcd*” monthly water balance model for the conterminous United States, *Water Resour. Res.*, *46*, W08507, doi:10.1029/2009WR008294.

1. Introduction

[2] Model-based projections of the impacts of hydroclimate variability and nonstationarity on water resources underpin policy-relevant decision making of major national and international significance [*Milly et al.*, 2008]. Continental scale water balance (WB) assessments can facilitate quantification of dynamic changes in water availability [*Arnell*, 1999], and help to communicate anticipated hydrologic impacts. While streamflow forecast systems at the local scale continue to improve (employing ever more sophisticated models), the simple WB model remains an important tool for characterization of hydrologic systems, and for identifying changes in hydroclimatic trends, systematic biases in data, and other potential problems. Similarly, while regional scale land surface models and global scale general circulation models will continue to improve (both in accuracy and detail), an improved understanding of local and regional scale WB dynamics can help in evaluating the implications and impacts of future change [*Rodell et al.*, 2004].

[3] With six decades of advances in hydrology, data acquisition, and computing technologies, we can now (in principle) represent the physicoclimatic conditions impacting water-related processes in considerable variety and detail. However, in any specific model application, there remains considerable ambiguity regarding which processes to include [*Sivapalan et al.*, 2003], making it difficult to produce reliable mappings of model-to-catchment characteristics [*McDonnell and Woods*, 2004]. As a community, we need to move toward broader and more general solutions rather than focusing on catchment specific solutions [*Andreassian et al.*, 2007]. In the context of this paper it is relevant that it is not at all clear whether existing conceptual representations are sufficient to represent the variety of natural catchment conditions observed in nature (climate, soils, topography, land cover, etc.), even for the relatively simple task of representing catchment-scale monthly water balances across the continental United States.

[4] Further, while our ability to perform complex model simulations while processing large volumes of data has improved dramatically, our ability to assimilate and evaluate model results has not kept pace. The process of model

[1] Department of Hydrology and Water Resources, University of Arizona, Tucson, Arizona, USA.

0043-1397/10/2009WR008294

building and diagnostic evaluation, analogous to the "learning" problem, is not trivial [*Gupta et al.*, 2008]. While some progress has been made on *quantitative* methods for evaluating the accuracy and precision of model predictions, *qualitative* methods for the evaluation of model consistency are poorly developed [*Sivapalan*, 2005] and remain weak at providing clear guidance for model improvement and correction.

[5] There is a clear need for well-conceived and systematic strategies for selecting model structures, establishing data requirements, estimating parameters, evaluating and tracking model performance, and diagnosing model deficiencies so that improvements can be made. These challenges can be better understood by first examining them in the context of simpler modeling situations.

[6] This paper therefore aims to understand the difficulties involved in WB model development. For the results to be generalizable, we emphasize the use of data from large numbers of catchments, and are interested in finding ways to automate the process. In particular, we examine the problem of monthly WB assessment across the continental United States, and begin with a prior conceptual hypothesis regarding dominant WB processes based on the popular "*abcd*" model [*Thomas*, 1981] used in several intercomparison studies [*Alley*, 1984; *Vandewiele et al.*, 1992; *Makhlouf and Michel*, 1994; *Mouelhi et al.*, 2006]. While this initial hypothesis may not be appropriate for all the locations tested, we are interested in whether it can be supported or contravened using the available data.

[7] For data, we exploit the monthly climate database prepared by *Vogel and Sankarasubramanian* [2005] from the Hydro-Climatic Data Network (HCDN) Streamflow data set compiled by the U.S. Geological Survey (USGS) [*Slack et al.*, 1993]. We use a subset of 764 stations, selected for comprehensive coverage of hydrogeological conditions and common time period. This large data set presents significant challenges to automation of the process, including (1) selection of model structures, (2) calibration and evaluation of model performance over hundreds of basins, and (3) strategies for making improvements to the model in a systematic way. With regard to this latter point, we seek to improve our understanding about how to formally diagnose the causes of model inadequacy thereby leading to their resolution [*Gupta et al.*, 2008].

[8] The goal outlined above is ambitious. This paper will focus, therefore, on the following questions.

[9] 1. How can one conduct a "robust" first-level evaluation of monthly WB model performance?

[10] 2. How well does the *abcd* model perform at the monthly time step over the continental United States?

[11] 3. What kinds of hydroclimatic and physical information can help in the classification of catchments and in explaining dominant processes and model performance at the monthly time step?

[12] 4. In which geographical regions is the HCDN data set capable of representing the dynamics of the monthly water balance?

[13] Section 2 of this paper provides background regarding water balance modeling, conceptual models and catchment classification for model selection. Section 3 contains a description of the methods and data used in the analysis. Sections 4–7 present the results of the diagnostic model evaluation. Section 8 discusses the results in the context of previous work. Sections 9 and 10 present and discuss our conclusions and recommend directions for future work.

2. Previous Work

[14] Water balance models (WBM) represent simple, yet refined, ways to understand hydrological processes and their interactions at the catchment scale. They can be used to anticipate changes in catchment behavior and to evaluate consequences of natural and/or human-induced changes to the system [*Wang et al.*, 2009]. Such models use precipitation and potential evapotranspiration or temperature as input, and generate estimates of catchment storage, actual evaporation and runoff.

[15] Since WBMs were first applied for agricultural planning [*Thornthwaite*, 1948], they have been used in the context of design and control of water resource systems [*Alley*, 1985; *Xu and Vandewiele*, 1995], understanding the regional hydrologic impacts of climatic change [*Gleick*, 1987; *Guo et al.*, 2002], and parameter regionalization studies [*Fernandez et al.*, 2000; *Vandewiele and Elias*, 1995; *Kling and Nachtnebel*, 2009], among others. *Vandewiele et al.* [1992] describe four broad types of applications: (1) to fill gaps in runoff data, (2) to estimate parameters for ungaged basins, (3) to generate estimates of soil moisture, and (4) to disaggregate model results to shorter time scales. *Xu and Singh* [1998] provide a comprehensive review.

[16] To test WBMs it is now common to use large numbers of catchments at national, and in some cases multinational, scales; examples include *Vandewiele et al.* [1992] (79 catchments), *Makhlouf and Michel* [1994] (91), *Hay and McCabe* [2002] (44), *Perrin et al.* [2001] (429), *Mouelhi et al.* [2006] (410), *Zhang et al.* [2008] (265), and *Perrin et al.* [2008] (900). In the United States, continental scale investigations include studies of long-term water balance [*Milly*, 1994; *Wolock and McCabe*, 1999; *Sankarasubramanian and Vogel*, 2002a], evaluation of land management scenarios and potential impacts of climate change [*Arnold et al.*, 1999; *Thomson et al.*, 2005], and generation of a widely used 50 year data set of water/energy fluxes at the 3 h time step [*Maurer et al.*, 2002]. Of direct relevance to our study, *Hay and McCabe* [2002] calibrated a model to 44 catchments and then extrapolated monthly WBM performance to 1646 catchments across the United States via a regression based on physical descriptors estimated from the calibrated catchments.

[17] In such studies, evaluation of model results is generally based on the mean squared error criterion (MSE) computed on the flows (or log or square root transformation of the flows), or its related normalization the Nash-Sutcliffe efficiency statistic (NSE) [*Nash and Sutcliffe*, 1970], along with some estimate of the bias. Calibration methods include stepwise parameter estimation [*Xiong and Guo*, 1999], simultaneous search [*Wang et al.*, 2009] using global algorithms such as SCE-UA [*Duan et al.*, 1992], and parameter set libraries based on a similarity measure [*Perrin et al.*, 2008]. While studies generally include some subjective level of diagnostic evaluation, it is not common to analyze the hydrogeoclimatic reasons for differences in model skill, or the reasons for poor performance. Further, it is not common to analyze the physical or hydroclimatic conditions under which particular model structures are appropriate. The typical goal of most studies is to justify a particular "new" model or family of models.

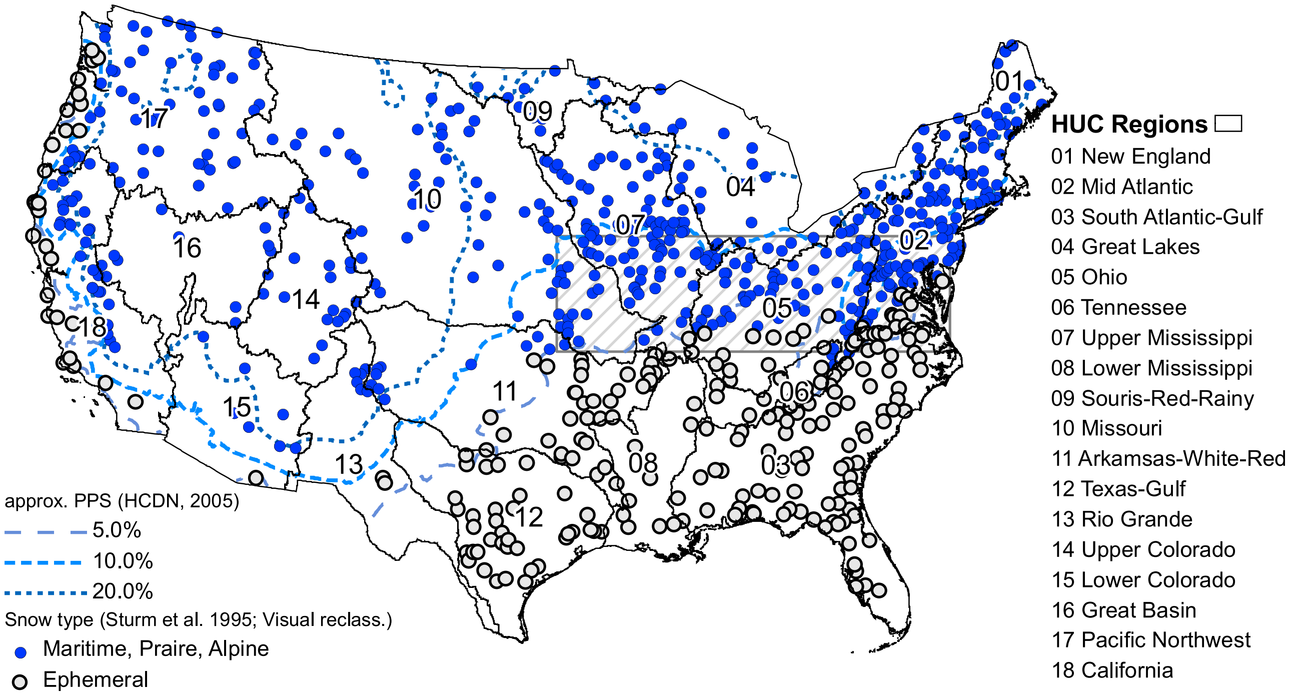

Figure 1. Locations of HCDN stations used in this analysis, classified as snow (solid circles) and no snow (open circles). HUC geographical regions level 2 are numbered from 01 to 18. Dashed lines indicate approximate contours of PPS. Hatched region indicates region where snow hypothesis was found to be ambiguous.

[18] Much of the catchment classification work is focused on issues of risk and water resources management [*Acreman and Sinclair*, 1986; *Wardrop et al.*, 2005] using statistical methods (clustering, principal components and discriminant analyses) applied to predominant physical characteristics (climate, land cover, soils, etc). Recently, the important role of classification for integrating our understanding of hydrologic systems has been recognized. *Wagener et al.* [2007] review existing approaches to catchment classification and suggest four essential requirements for a general scheme. *Berger and Entekhabi* [2001], *Yadav et al.* [2007] and *Bai et al.* [2009] use similarity indices and hydrologic signatures to discriminate catchments based on dominant behaviors, functions and processes. *Dunne* [1983] presents a diagram connecting runoff processes to climate, land cover, topography and soils control. *Wolock and McCabe* [1999] study the principal aspects explaining spatial variability in mean annual runoff for the 344 climate divisions in the conterminous United States. *Berger and Entekhabi* [2001] and *Sankarasubramanian and Vogel* [2002b] develop regression relationships to estimate runoff ratios from catchment descriptors. *Winter* [2001] introduces the concept of hydrologic landscape regions (HLR), and classifies the United States landscape in terms of land surface slope, hydraulic properties of soils, geological setting, and difference between precipitation and evapotranspiration.

[19] Data sets have been created to study aspects of catchment function related to climate and other physical variables [*Smith et al.*, 2004; *Duan et al.*, 2006]. In the United States, the principal mechanisms for cataloging river basin information are the hydrologic unit maps and hydrologic unit codes (HUCs) (Figure 1), which organize catchments in terms of geographic location. *Lins* [1997] points out that, although an important tool for catchment management, the HUC classification is based only on *topographical* information and does not capture other important aspects of system hydrology such as patterns of streamflow.

[20] In summary, while considerable work has been done to evaluate monthly WBMs for various locations and under specific conditions, no extensive analysis of model performance (and associated structures and parameters) for different hydroclimatic and physiographic conditions has been reported. At the continental scale, model evaluation has typically been done only at the outlets of large catchments, or streamflow estimates reported on flow grids at the annual time scale. Further, classification systems based on catchment characteristics and/or dominant processes are rarely used to aid in the specification of model structures. With this paper, we hope to encourage more rigorous evaluation and diagnosis of hydrologic models, using the large catchment data sets available in the United States to evaluate and discuss model hypotheses and identification methods, and to identify locations where existing model representations are not adequate.

3. Methods and Data Sets

3.1. Model Identification

[21] Model identification involves a recursive set of steps including (1) selection of study sites and data, (2) selection of a model hypothesis to be tested, (3) calibration of the model, (4) evaluation of model performance and assessment of adequacy, (5) modification of the model hypothesis, and (6) reiteration of these steps until satisfied. In practice, more

than a single iteration is rarely achieved. Further, the model evaluation process remains diagnostically weak, with little guidance toward productive directions for model improvement [*Gupta et al.*, 2008]. Model correction therefore depends highly on the creativity of the practitioner and remains more of an art than a science [*Savenije*, 2008].

[22] One of our goals is to understand how to diagnose causes of model inadequacy, leading to their resolution. Therefore we will adopt the following general procedure.

[23] 1. Begin with a "simple" problem.

[24] 2. Establish an informative data set D.

[25] 3. Select a conceptual model structure (hypothesis H = C).

[26] 4. Select a mathematical model structure (hypothesis H = M|C) consisting of a set of model equations M consistent with the conceptual model C.

[27] 5. Select model parameters θ that "calibrate" the model to the data (parameters θ|H,D), by maximizing the likelihood L(D|H,θ) that the data could have been generated by the model.

[28] 6. Evaluate the suitability of the model (examine H|D).

[29] 7. Diagnose model problems.

[30] 8. Propose model corrections.

[31] 9. Return to one of the earlier steps as appropriate and iterate.

[32] 10. Repeat entire procedure on a more complex problem.

[33] This procedure can be generalized to handle multiple conceptual and mathematical model structures examined at the same time [*Neuman*, 2003; *Clark et al.*, 2008], and issues of space and time scale [*Blöschl*, 2006]. Here, the problem is monthly water balance modeling for the continental United States using an available data set of 764 catchments. We use several physical and climatic descriptors to explore the causes of good and bad model performance. Our initial conceptual model and mathematical model structure is based on the *abcd* model applied in lumped fashion at the catchment scale and monthly time step; note that issues of spatially distributed modeling and finer time scales (while arguably important) are not examined here. Within this context, we investigate the process of model performance evaluation from a diagnostic perspective, leading to a proposal for modifications to the model hypothesis and improvements in the procedure used for model calibration. We also explore relationships that seek to explain model performance and parameter estimates based on the descriptor variables.

[34] Due to the complexity of the analysis, and sheer volume of results to be analyzed and presented, this paper will report only on the first iteration through the model identification loop presented above. However, we go further than previous studies in that we establish a clear basis for the next iteration through the loop, the results of which will be reported in follow-up publications.

3.2. Data Sets

[35] *Vogel and Sankarasubramanian* [2005] compiled a hydroclimatological data set of monthly streamflow, precipitation, potential evapotranspiration, and average daily minimum and maximum temperatures for 1376 catchments in the United States using catchment boundaries derived from a 30 arc sec DEM. The streamflow data was compiled by the USGS [*Slack et al.*, 1993] to be representative of the climatology of the continental United States while being relatively unaffected by human influences such as streamflow regulation, impacts of groundwater pumping, or changes in measurement device. Lumped average values of precipitation and average minimum and maximum daily temperature data were computed by the Precipitation Elevation Regression on Independent Slopes Model (PRISM) climate analysis system; *Daly et al.* [1994]), interpolated from point measurements onto a 2.5 min (~4 km) grid while correcting for topography (elevation) and other factors. Monthly potential evaporation was estimated using the *Hargreaves and Samani* [1982] method.

[36] In this paper, we use a subset of 764 catchments (Figure 1) for which data spanning a common 40 year time period (water years 1951–1990 inclusive) are available. The log areas of the catchments are approximately normally distributed around a median of 1200 km^2, with 95% of the catchments being between 200 and 15,000 km^2. The lowest 5 percentile of catchments (areas <200 km^2) are located mainly in the New England region (HUCs 01 and 02), mountains of California (HUC 18), and headwaters of the Colorado and Rio Grande River basins. The largest 5% of catchments (areas >15,000 km^2) are located mainly across the Northern Plains and the South Atlantic–Gulf/Texas-Gulf regions; the two very largest catchments (222,000 and 308,000 km^2) are in the Upper Mississippi basin. Other data used in computing catchment descriptors include HCDN auxiliary tables of catchment area, hydroclimatic variables, topographic parameters, soil properties and variables related to temperature and energy balance [*Vogel and Sankarasubramanian*, 2005], the USGS hydrologic units map [*Seaber et al.*, 1987], and the hydrologic landscape regions HLR map [*Wolock et al.*, 2004].

3.3. Catchment Descriptors

[37] General catchment descriptors in the HCDN database include indices derived from time series of precipitation (P), streamflow (Q), potential evapotranspiration (PE), and temperature (T). These "descriptor variables" include (1) major indices of hydroclimatology, including runoff ratio (Q/P), aridity index (PE/P), evapotranspiration ratio ((P – Q)/P = E/P), and mean annual temperature (T_a); (2) major physiographic indices, including catchment area, average elevation and slope derived from a 1 km digital elevation map (DEM); (3) variables related to snow dynamics, including average temperatures for different times of the year, average snow water equivalent (SWE), average aspect, percentage of precipitation as snow (PPS), and heating and cooling degree days; and (4) streamflow indices, including base flow index (BFI) and base flow normalized by precipitation (BF/P) [*Institute of Hydrology* 1980], and flow duration curve indices such as the flow magnitude for the lowest 1st percentile (Q_{1p}).

[38] In the context of this work, the catchment descriptors we use are mainly mean areal averages. Of course, such average descriptors may not adequately represent the diverse characteristics of large catchments, particularly with regard to physical properties and snow related information. In this study, we apply this simple aggregate approach with a view to determining whether there are area thresholds above

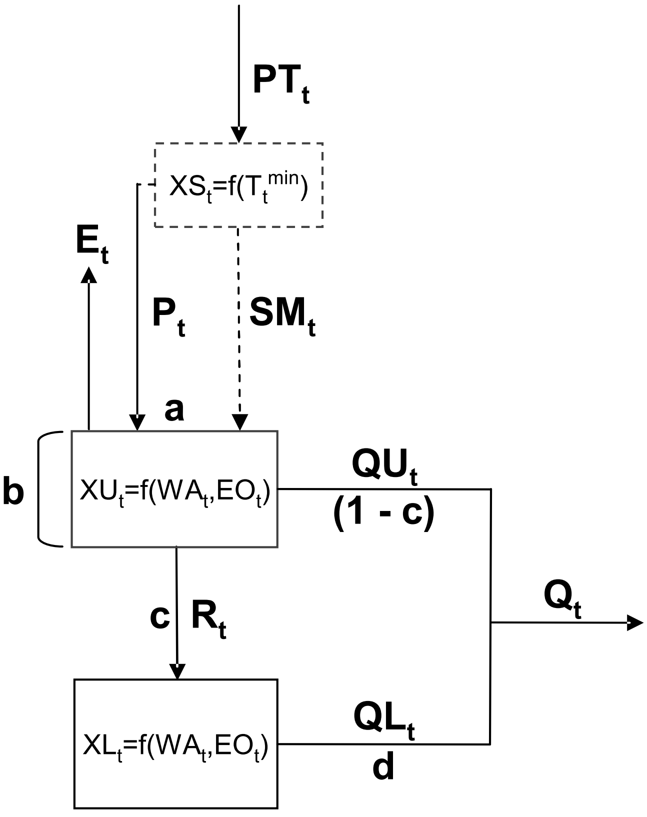

Figure 2. Structure of the *abcd* model, with snow component added (dashed lines).

which the lumped approach is inadequate, rather than imposing some arbitrary threshold a priori.

3.4. Model Hypothesis

[39] We select the *abcd* model formulation [*Thomas*, 1981] to represent our initial WBM hypothesis for regions where snow accumulation and melt is not a significant component of the monthly water balance. The model is based on the *Thornthwaite* [1948] conceptual framework for parsimonious WB modeling, but incorporates a more realistic representation of infiltration by allowing streamflow to occur even under conditions of moisture deficit (Figure 2). It applies the continuity equation to a control volume representing the upper soil zone, from which evapotranspiration is assumed to occur, so that

$$P_t - E_t - R_t - QU_t = \Delta XU = XU_t - XU_{t-1} \qquad (1)$$

where P_t is total precipitation for the month, E_t is actual evapotranspiration, R_t is recharge to groundwater storage, QU_t is upper zone contribution to runoff, and XU_t and XU_{t-1} represent upper soil zone soil moisture storage at the current and previous time steps respectively. Equation 1 can be rearranged to obtain $(P + XU_{t-1}) = (E_t + XU_t) + QU_t + R_t$, which facilitates definition of the two main state variables of the model, "available water" ($WA_t = P_t + XU_{t-1}$) and "evapotranspiration opportunity" ($EO_t = E_t + XU_t$) [*Sankarasubramanian and Vogel*, 2002a]. A key assumption is that EO_t is nonlinearly related to WA_t in a manner such that EO_t increases quickly with available water for $WA_t < b$ (water limited conditions), but asymptotically approaches a maximum constant value of b for $W_t \gg b$ (energy limited conditions):

$$EO_t(WA_t) = \frac{WA_t + b}{2a} - \sqrt{\left(\frac{WA_t + b}{2a}\right)^2 - \frac{WA_t \cdot b}{a}} \qquad (2)$$

in a manner analogous to the Budyko relationship [*Budyko*, 1974]. The parameter a controls the concavity of the relationship between WA_t and EO_t and emulates the "propensity in many catchments for runoff to occur well before the soils are saturated to capacity" [*Thomas*, 1981, p. 23]. The parameter b represents the water holding capacity (proportional to depth) of the upper soil zone.

[40] Evapotranspiration opportunity EO_t is further partitioned into actual evapotranspiration E_t and residual soil moisture storage XU_t by relating the rate of soil moisture loss to potential evapotranspiration, leading to the nonlinear relationship $E_t = EO_t \cdot \{1 - \exp(-PE_t/b)\}$. Water available for runoff ($WA_t - EO_t$) is further partitioned into upper zone contribution to runoff QU_t and recharge to groundwater R_t by the parameter c, according to $QU_t = (1 - c) \cdot (WA_t - EO_t)$ and $R_t = c \cdot (WA_t - EO_t)$. Recharge R_t is added to the lower soil zone state variable XL_{t-1} and base flow to the stream is computed according to the linear recession relationship $QL_t = d \cdot (XL_t)$. Using continuity, we update $XL_t = (XL_{t-1} + R_t) \cdot (1 + d)^{-1}$. Finally total streamflow is computed as $Q_t = QU_t + QL_t$.

[41] For regions dominated by snow dynamics, we construct an augmented *abcd*-snow model by including a simple temperature-based snow accumulation and melt component (Figure 2, dashed lines), so that total precipitation PT_t is partitioned into precipitation as rain PR_t and precipitation as snow PS_t according to

$$PS_t = \begin{cases} 0 & T^{rain} < T_t^{\min} \\ PT_t \cdot \dfrac{T^{rain} - T_t^{\min}}{T^{rain} - T^{snow}} & T^{snow} \leq T_t^{\min} \leq T^{rain} \\ PT_t & T_t^{\min} < T^{snow} \end{cases} \qquad (3)$$

where $T_t^{\min}$ is the average daily minimum temperature for the month, T^{rain} is the temperature threshold above which all the precipitation falls as liquid water, and T^{snow} is the temperature threshold below which all the precipitation falls as snow. Accumulated snow water equivalent, represented by the state variable XS_t, is stored in an infinite capacity tank which releases snowmelt SM_t at a rate determined by the melt coefficient m and the temperature thresholds according to

$$SM_t = \begin{cases} 0 & T_t^{\min} < T^{snow} \\ (XS_{t-1} + PS_t) \cdot m \cdot \dfrac{T^{rain} - T_t^{\min}}{T^{rain} - T^{snow}} & T^{snow} \leq T_t^{\min} \leq T^{rain} \\ (XS_{t-1} + PS_t) \cdot m & T^{rain} < T_t^{\min} \end{cases} \qquad (4)$$

[42] This WBM formulation does not model snow sublimation or other, more complex, spatiotemporal dynamics of the snow accumulation/ablation process. Precipitation multipliers were not used since we assume that the PRISM approach has largely corrected for precipitation measurement biases [*Daly et al.*, 1994]. Further, we assume that the effects of submonthly distribution of timing and intensity of precipitation events, potential evapotranspiration and temperature variations, and other factors are negligible. *McCabe*

and Wolock [1999] and *Hay and McCabe* [2002] have used similar approaches.

3.5. Snow and No-Snow Classification of Catchments

[43] Our model hypothesis requires that the catchments be classified as "snow" or "no snow" based on whether snow accumulation/melt dynamics are believed to play an important role. Several approaches were tested, based on an evaluation of winter season average temperatures, variation of runoff ratios at the end of winter, estimates of PPS reported in the HCDN database (see Figure 1), remote sensing information, and DEM and catchment boundaries. Via experimentation, we settled on the following two-step procedure. (1) First classify the catchments with regard to predominant snow cover type using the raster map data set prepared by *Sturm et al.* [1995]. (2) Next manually correct the classification of catchments in transition zones. For the latter step, a visual inspection of GIS information was conducted to check locations of catchment boundaries and to examine factors such as elevation, headwater location, PPS and winter average temperature. No-snow catchments having headwater areas in snow regions or at high elevations close to snow regions were reclassified as snow.

[44] Based on this, 241 catchments are classified as no snow (meaning ephemeral snow, for which the four parameters (a, b, c, d) and two initial states (XU_0, XL_0) of the *abcd* model must be estimated. The remaining 523 catchments are classified as snow for which the additional three parameters (T^{rain}, T^{Δ}, m) of the augmented *abcd*-snow model must be estimated, where $T^{\Delta} = T^{rain} - T^{snow}$ represents the width of the rain-snow transition interval. Because the simulation period begins on 1 October, the initial snow accumulation XS_o is assumed to be zero; while this assumption may not be strictly correct for some locations, a check showed little sensitivity to the assumption. While significant consistency was observed in the variables used for snow to no-snow classification, it was not find to be possible to achieve the partitioning in a simple manner using a single set of descriptors. In general, we observed the following.

[45] 1. 96% of no-snow catchments occur at elevations <1000 m, while 45% of the snow catchments have mean elevation >500 m and 28% are higher than 1000 m.

[46] 2. 67% of snow catchments have PPS > 10%, while 80% of no-snow catchments have PPS < 5% (with only 3 catchments above 10%).

[47] 3. 67% of no-snow catchments have average max temperature >10°C in winter (December, January, February), while only 1% of snow catchments are above this threshold.

[48] 4. Runoff ratios are larger for snow catchments; 29% of snow catchments have Q/P > 0.5 compared to 11% for no-snow catchments.

[49] 5. Other variables, including catchment area and aridity index, did not exhibit snow to no-snow discriminator power.

3.6. Calibration Approach

[50] To characterize the hydroclimatic locations and conditions under which the selected model hypotheses are (and are not) adequate, we calibrate the appropriate model structure to match the 1951–1960 10 year period input-output behavior at each of the 764 catchments, via conventional single-criterion optimization (parameter estimation) using the shuffled complex evolution (SCE-UA) [*Duan et al.*, 1992] algorithm to minimize the MSE criterion. A preliminary analysis testing different criteria (including mean absolute error and log transformations of the flows) on a subset of catchments showed the MSE to give generally good results at the monthly time step when evaluated via visual examination of hydrographs and various measures including overall streamflow bias [*Martinez*, 2007]. The use of MSE is consistent with the work of *Thomas* [1981], *Alley* [1984], *Vandewiele et al.* [1992] and *Makhlouf and Michel* [1994]. Feasible parameters ranges were specified as follows. Parameters a, c, d and m (dimensionless) were allowed to vary on [0, 1]. Parameter b (mm), and initial states XU_0 and XL_0 were allowed to vary on [0, 4000], to avoid hitting the bounds. Parameter T^{rain} (°C) was allowed to vary on [−10 10] and parameter T^{Δ} (°C) was allowed to vary on [0 to 100]. These large ranges were used because the catchment average representativeness of PRISM temperature estimates varies with catchment area [*Daly et al.*, 1994].

4. "Classical" Evaluation of Model Performance

[51] For an initial evaluation of model calibration results we examine the NSE statistic of similarity between simulated and observed flows (NSE = 1 − MSE/σ_o^2 where σ_o^2 is variance of the observed monthly flows). Although the NSE has inherent limitations [*Schaefli and Gupta* 2007], it is the criterion most commonly used and therefore allows relative comparison with other studies. Figure 3 shows a clear correspondence between calibration period NSE and geographical characteristics. The PPS contours (dashed lines) were used as guidance in the snow versus no-snow classification. The cumulative distribution function (cdf) of NSE (Figure 3, inset) shows the range of model performance achieved, and is used to categorize performance into four groups, namely, "good," "acceptable," "poor" and "bad." An examination of separate NSE cdf's for snow and no-snow basins (not shown) shows rapid deterioration in performance below threshold values of 0.63 and 0.67 respectively. We therefore subjective select the (conceptually reasonable) threshold values of 0.75, 0.67 and 0.59 to partition catchments into good (1.0–0.75), acceptable (0.75–0.67), poor (0.67–0.59), and bad (<0.59) model performance, respectively. Consequently (Table 1), shows that 92% of the no-snow and 81% of the snow catchments are classified as acceptable to good; these are located predominantly in the wetter eastern United States and western coast. Meanwhile 8% of the no-snow and 19% of the snow catchments are classified as poor to bad; these tend to occur more in the drier central regions. Interestingly, model performance in the southeastern United States decreases with proximity to the ocean, and also as we move from humid toward subhumid and arid regions.

[52] Model performance also shows strong correspondence with major hydroclimatic controls. A decision tree analysis indicates that the E/P ratio (indicating wetness) and the mean annual temperature T_a have the greatest power to explain differences in NSE performance (see Appendix A for details). Interestingly, nonhydroclimatic variables do not help to explain differences in model performance. In fact we see only slight deterioration in calibration period NSE with catchment area: 90% of the smallest catchments (area <200 km^2), 80% of the intermediate catchments, and 66% of

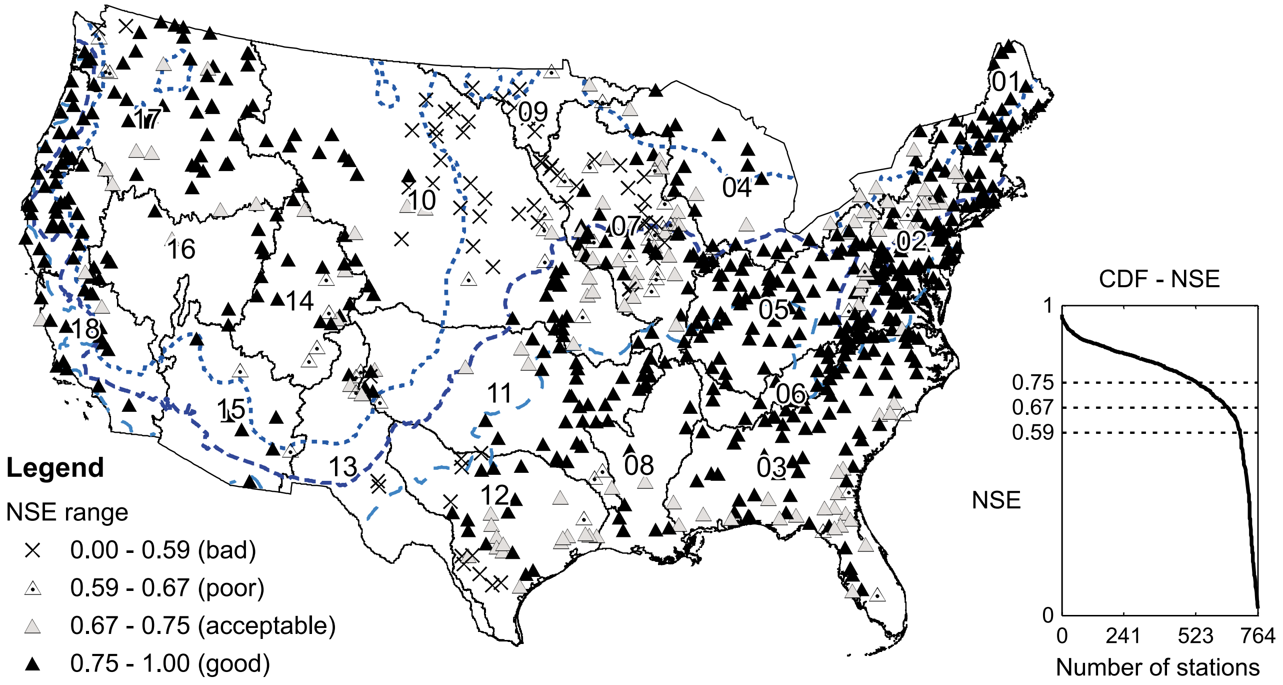

Figure 3. Map showing spatial distribution of NSE performance for the calibrated WB model hypothesis. Dashed lines show approximate PPS contours. Inset shows the cumulative distribution function of NSE performance, indicating thresholds used for the performance classification.

the largest catchments (area >15,000 km^2) including the four largest catchments, have performance in the good to acceptable range. In general, worst performance occurs for water limited catchments having a high *E/P* ratio.

[53] In testing the model hypothesis on snow catchments, we calibrate both *abcd* and *abcd*-snow model structures, expecting to see a statistically significant difference in *performance* for catchments where snow dynamics constitute a significant part of the water balance. Figure 4 compares NSE performance with and without the snow component. For several catchments, corresponding mainly to the Rocky Mountains in the west and New England in the northeast, NSE improves dramatically from NSE ~0.1–0.3 to NSE ~0.7–0.9, supporting the use of the augmented model. However, 50% of the catchments show less than 0.05 improvement in performance, and 28% show improvements smaller than 0.01. The latter are generally located in the east-central region, constituting the buffer zone between catchments classified as snow and no-snow (hatched region in Figure 1). This result suggests that the latter catchments should be reclassified as no snow, or (as suggested by a reviewer), that our classification into only two categories (snow versus no snow) is not sufficient, since catchments with consistent winter snow cover area may be easier to model than those with variable cover. In support of this, an examination of the *National Climatic Data Center* [2002] atlas revealed that catchments with largest improvement in performance are those having greater than 56 d/yr with

Table 1. Variables Found to Be Important in the Discrimination of NSE Classes Using Decision Trees

	No Snow Catchments				
Catchments	General Location	$E/P \geq 0.96$	$E/P < 0.96$ and $T_a < 16°C$	$E/P < 0.96$ and $T_a \geq 16°C$	Total
Good (1.00–0.75)	Predominant along NW coast; Southeast regions	4	129	47	180
Acceptable (0.67–0.75)	Scattered along Southeast coast	0	5	37	42
Poor (0.59–0.67)	Scattered	0	0	7	7
Bad (<0.59)	Mainly central HUC 12,13	9	1	2	12
Total	-	13	135	93	241
	Snow Catchments				
Catchments	General Location	$E/P \geq 0.88$ and $T_a < 5°C$	$E/P \geq 0.88$ and $T_a > 5°C$	$E/P < 0.88$	Total
Good (1.00–0.75)	Eastern HUC 01; scattered through HUC 17, 14	7	6	336	349
Acceptable (0.67–0.75)	Scattered, mainly Northern Plains and HUC 02	8	2	65	75
Poor (0.59–0.67)	Scattered, mainly Northern Plains	3	5	35	43
Bad (<0.59)	Northern Plains HUC 07, 09 10	39	0	17	56
Total	-	57	13	453	523

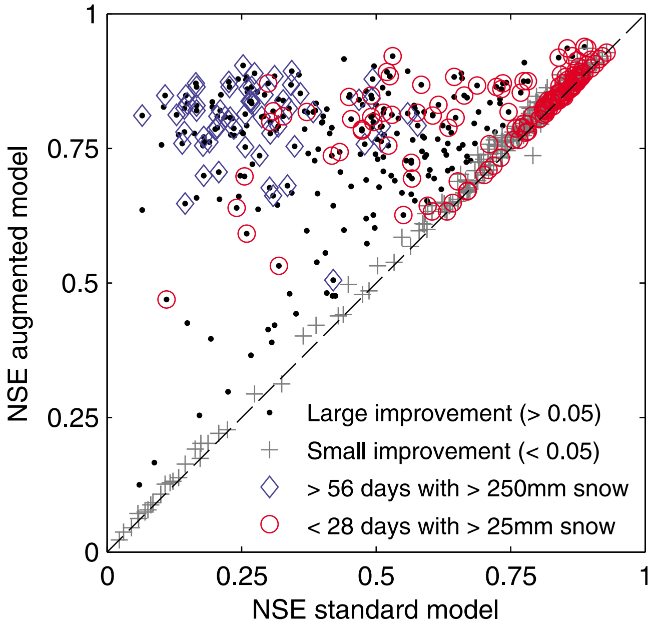

Figure 4. Comparison of NSE, before and after adding the snow component, for the 523 catchments classified as snow.

daily average snow depths larger than 250 mm, while catchments with marginal performance improvements have very few snow days and relatively small average snow depths.

5. "Diagnostic" Evaluation of Model Performance

[54] The assessment in section 4, while useful to isolate problems in calibration methodology and to identify regions having different levels of performance, is limited by weaknesses inherent in use of NSE as a performance measure. *Gupta et al.* [2009] show that NSE decomposes into three terms: bias error, variance error and linear correlation, which measure different aspects of model performance:

$$\mathrm{NSE} = 2\alpha r - \alpha^2 - \beta_n^2 \tag{5}$$

[55] Here α is the ratio of standard deviations ($\alpha = \sigma_s/\sigma_o$) of the simulated and observed quantities being compared, β_n is the long-term simulation bias normalized by the observed standard deviation ($\beta_n = (\mu_s - \mu_o)/\sigma_o$), and r is the linear correlation coefficient. When optimizing on MSE (or NSE), interdependence among these components can cause the model to exhibit significant bias and variance errors. Further, although the optimal value for $\beta_n = 0$, the optimal value for the standard deviation ratio is $\alpha = r$ which is always less than 1.0, so that the "MSE-optimal" model will tend to underestimate the variability of the flows.

[56] Substituting $\beta_n = 0$ and $\alpha = r$ into equation (5) we get NSE ~ r^2, a fact which can be verified by plotting the results for our 764 catchments (Figure 5a); the calibrated points fall very close to the NSE = r^2 line. However, Figures 5b and 5c show that even catchments with high NSE can have significant errors in reproduction of first and second moment statistics (μ_o and σ_o) of the flows, indicating that NSE after optimization is dominated by r while being relatively insensitive to errors in reproduction of the long-term mean and variability of flow. Further, since $\alpha \rightarrow r$ during calibration with *NSE*, there is a strong tendency to underestimate the observed flow variability – this is true even if the model does have the capability to simulate the observed variability of flows (of course, in some catchments it may not be able to do so due to severe model structural or input errors). Since the goal of WB modeling is to simulate the dynamics of catchment water balance, the model should be capable of generating accurate estimates of mean and variability of flow (unfortunately, we do not have data to test other fluxes). Therefore, we next look more closely at the distribution of model performance obtained, and reclassify the results based on their success in reproducing the long-term mean and standard deviation of monthly flow.

[57] First we assess model performance in terms of water balance error $\Delta\mu = (\mu_s - \mu_o)$. Because this can be assessed either in terms of fraction of streamflow ($\Delta\mu_q/\mu_q$) or fraction of precipitation ($\Delta\mu_q/\mu_p$), we examine both quantities. Figure 6a plots $\Delta\mu_q/\mu_q$ against $\Delta\mu_q/\mu_p$, showing, as expected, that the quantities are strongly correlated. Selecting 5% and 10% as performance thresholds on $\Delta\mu_q/\mu_q$ (streamflow), corresponding to 2% and 4% thresholds on $\Delta\mu_q/\mu_p$ (total precipitation), we group the calibrated catchments into three bias error classes (Table 2), good (71%), acceptable (17%) and bad (11%). For no-snow catchments, the catchment descriptors explaining differences in water balance performance are *Q/P* (runoff ratio) and *PE/P* (aridity ratio). Catchments with large positive bias tend to be in arid regions

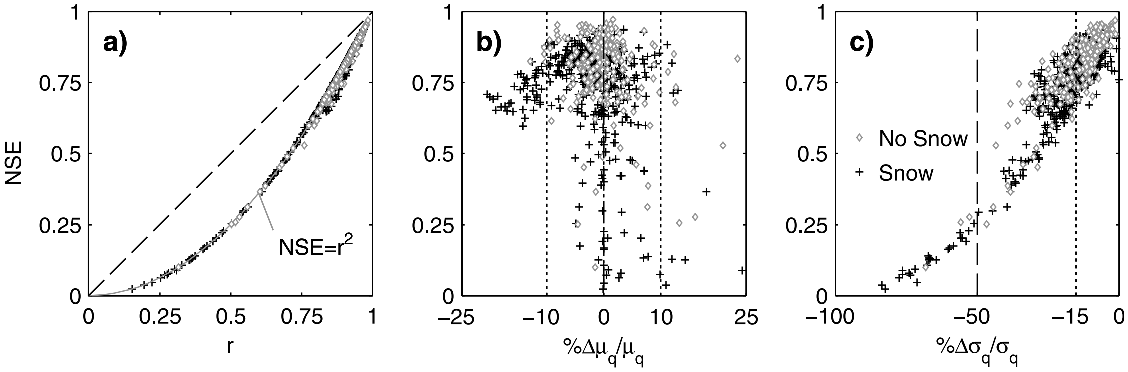

Figure 5. Comparison of NSE value to three diagnostic components (a) r, (b) $\%\Delta\mu_q/\mu_q$, and (c) $\%\Delta\sigma_q/\sigma_\mathrm{p}$ for no-snow (pluses) and snow (diamonds) catchments.

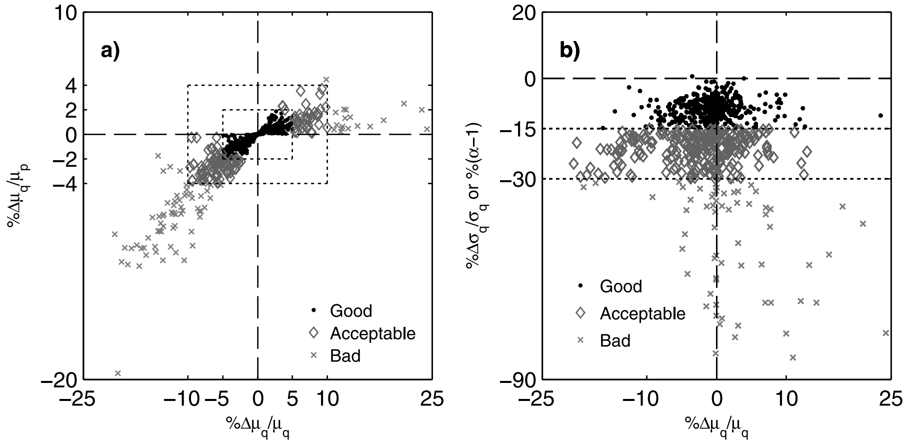

Figure 6. Classification of catchments based (a) on bias and (b) variability errors. Note that $\Delta\sigma_q/\sigma_p$ thresholds of −30% and −15% correspond to α values of 0.70 and 0.85, respectively.

(*Q/P* < 0.25 and *PE/P* > 1.4) scattered mainly along the central United States humid-arid transition zone, while catchments with large negative bias tend to be in humid regions (high *Q/P* and low *PE/P*), predominantly in the Appalachian Mountains. For snow catchments the behavior is similar; those with negative bias are located along the Appalachian Mountains and those with positive bias are throughout the central United States. Additionally, variables controlling snow storage such as annual temperature and elevation appear to have some discriminatory power.

[58] Next we assess model performance in terms of reproduction of flow variability defined as $\%\Delta\sigma_q/\sigma_q$ (Figure 6b); note that $\Delta\sigma_q/\sigma_q = (\alpha - 1)$. Setting 15% and 30% as performance thresholds on $\%\Delta\sigma_q/\sigma_q$ ($\alpha = 0.85$ and 0.70), we group catchments into three variability error classes, good (40%), acceptable (52%) and bad (8%). As predicted by theory, all but one of the calibrations obtained by maximizing NSE have negative $\%\Delta\sigma_q/\sigma_q$, meaning that they underestimate the variability of the flow distribution. The most significant variability errors occur for catchments in the Northern Plains and the western Texas/Gulf region. In general, variability error is worse for drier regions (*Q/P* < 0.15 and PE*/P* > 1.4).

[59] Combining the discrimination of catchments based on NSE, $\%\Delta\mu_q/\mu_q$ and $\%\Delta\sigma_q/\sigma_q$, we reclassify the 764 catchments (Figure 7) into those having good performance on *all* three measures (44%), bad performance on *any* of the three measures (25%), and all others classified as acceptable (31%). Good performance tends to be located in the lower Mid-Atlantic region (HUC 02), Ohio (HUC 05), upper South Atlantic–Gulf (HUC 03), and for catchments scattered along the west coast and northern Rocky Mountains. Bad performance tends to be located along the Appalachian Mountain divide (especially northern portion of HUC 02), the northern plains, the central United States, the northern Rio Grande (HUC 03) and Upper Colorado, and the Sierra Nevada and Pacific Northwest regions of the West Coast. Acceptable performance is scattered throughout the United States.

[60] For no-snow catchments, the descriptor having greatest power to explain differences in overall calibration

Table 2. Variables Found to Be Important in the Discrimination of Flow Bias Classes[a]

	Predominant HUC[b]	PE/*P* ≥ 1.4 and *Q*/*P* ≥ 0.25	PE/*P* ≥ 1.4 and *Q*/*P* < 0.25	PE/*P* < 1.4 and *Q*/*P* ≥ 0.25	PE/*P* < 1.4 and *Q*/*P* < 0.25	Total
			No Snow			
Bad (+)	-	0	8	0	0	8
Acceptable (+)	-	1	9	3	2	15
Good (+)	Scattered	3	15	58	8	84
Good (−)	Scattered	2	12	82	10	106
Acceptable (−)	5,6	1	1	16	2	20
Bad (−)	6	0	0	8	0	8
Total	-	7	45	167	22	241
			Snow			
Bad (+)	-	0	9	1	0	10
Acceptable (+)	-	2	16	10	3	31
Good (+)	Scattered	16	42	53	23	134
Good (−)	Scattered	18	30	162	11	221
Acceptable (−)	1,14	10	3	53	1	67
Bad (−)	4,5	1	4	55	0	60
Total	-	47	104	334	38	523

[a] Here (+) and (−) indicate positive and negative bias, respectively.

[b] HUC with more than 10 watersheds and more than 50% for the respective class.

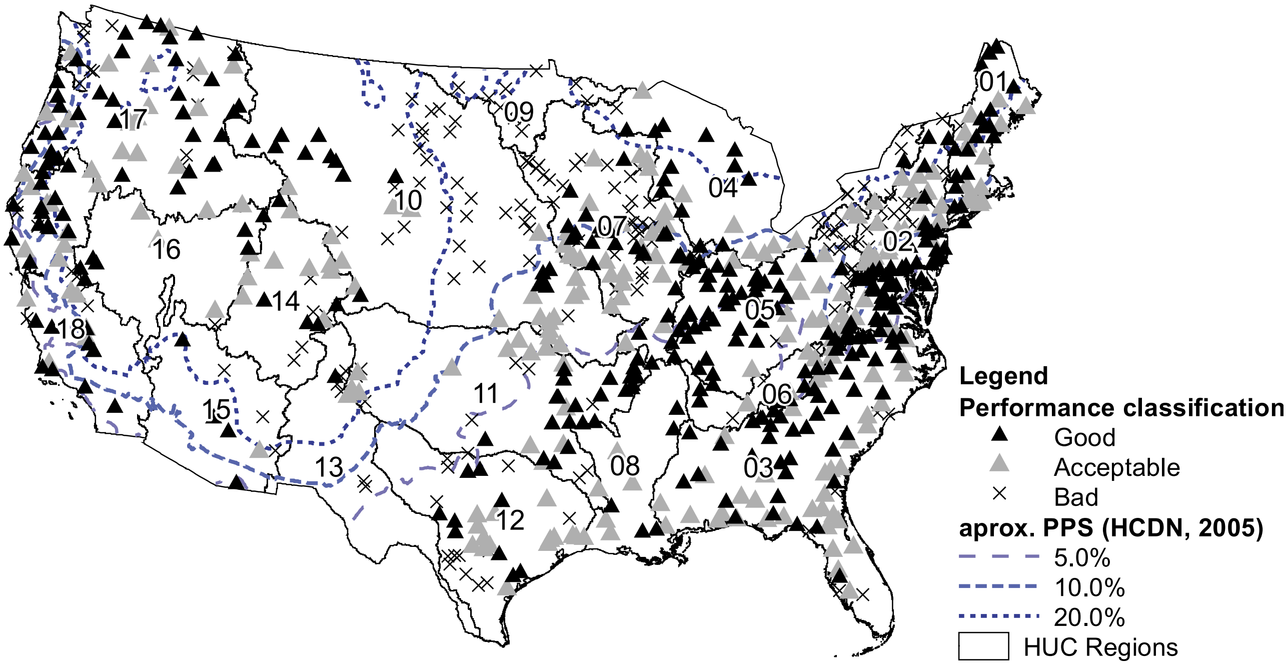

Figure 7. Spatial distribution of calibration period model performance classified into good, acceptable, and bad.

performance is *Q/P*. In general, catchments having intermediate values of runoff ratio (0.27 < *Q/P* < 0.75) have good performance, while catchments with extreme values for *Q/P* have bad performance. For catchments with 0.02 < *Q/P* < 0.27, the catchment area and magnitude of lowest flows also show some discriminatory power. For example, in this interval, performance is good for all catchments with areas less than 500 km^2 while bad performance is more common for catchments having low flow magnitude for the lowest 1st percentile (Q_{1p}).

[61] For snow catchments we were unable to find a unique set of catchment descriptors that explain overall calibration performance. Therefore we conducted the analysis separately for three distinct geographical regions. For the mountainous western United States (HUC 13 to 18), catchments with higher mean annual temperature (T_a > 6°C) have good performance, while bad performance is common for catchments with low mean annual temperature (T_a < 6°C) and high runoff ratio (*Q/P* > 0.60). It is possible that the large *Q/P* observed at high elevations may be indicative of underestimation of precipitation (a violation of our study assumption). These stations are also characterized by high snow water equivalents, low winter temperatures, and large low magnitude flows (Q_{1p}). Note that *Daly et al.* [1994] warn that PRISM interpolations of precipitation and temperature may have limitations in this region.

[62] For the arid north central United States (HUC 07, 09, 10), bad performance occurs for relatively arid catchments with low runoff ratio (*Q/P* < 0.14), high evapotranspiration ratio (*E/P* > 0.85), small low flow magnitude (Q_{1p}), and low *PPS* (10–20%). For this region, *Hughes and Robinson* [1996] report high interannual variation of precipitation and snow. It is also well known to have a high percentage of irrigated cropland and bare ground land cover, and the presence of potholes in the northern prairies [*Euliss et al.*, 1999]. The poor performance could be due to many causes including snow sublimation, and groundwater and evapotranspiration loss. Further, an artificial neural network analysis [*Martinez*, 2007] suggests inadequate information content of the available data.

[63] For the eastern United States (HUC 01, 02, 04, 05) bad performance occurs for catchments with medium to high runoff ratios (*Q/P* > 0.46) and higher mean annual temperatures (T_a > 6°C). In particular, bad performance is concentrated on the higher elevations of the western face of the Appalachian Mountains, while good performance occurs in the New England region (HUC 01) where winter temperatures and elevations are lower.

[64] In general, the major controls on model performance appear to be hydroclimatic variability and the intermittence of water in the catchments, explained using indices such as runoff ratio (*Q/P*) and aridity ratio (PE/*P*) and magnitude of 1st percentile flows (Q_{1p}). For snow regions, regional discrimination is related to importance of snow storage and variability of snow accumulation, explained using indices such as annual temperature (T_a). However, geographical and geomorphic controls such as wetlands and potholes, and indices of intramonthly variability, also appear to be important, but were not represented in the list of catchment descriptors available for this analysis.

6. Evaluation of Model Parameters

6.1. Evaluation of Soil Zone Parameters *a*, *b*, *c*, and *d*

[65] To evaluate properties of the calibrated model parameters, we restrict our attention to the 127 no-snow and 212 snow (total = 339) catchments where good model performance is achieved, because for the remaining catchments the parameter estimates may either be biased or not meaningful. In general, the estimated parameter values are found to fall comfortably within expected ranges reported in the

Table 3. Comparison of the *abcd* Parameters Ranges With Those of *Alley* [1984] and *Vandewiele et al.* [1992][a]

Study	*Alley* [1984]	*Vandewiele et al.* [1992]	This Study, No Snow	This Study, Snow
Number of catchments	10	79	127	212
Statistics	Mean, CV, [Range]	Mean, [Range]	Mean, CV, [Range]	Mean, CV, [Range]
Parameter *a*	0.992, 0.007, [0.975–0.999]	0.986, [0.96–0.999]	0.977, 0.019, [0.873–0.999]	0.933, 0.17, [0.04–1.0]
Parameter *b*	30, 0.35, [14–50]	475, [260–1900]	393, 0.36, [133–922]	401, 0.81, [98–1590][b]
Parameter *c*	0.16, 1.0, [0.01–0.46]	0.270, [0.04–0.70]	0.229, 1.02, [0–1]	0.21, 1.23, [0–1]
Parameter *d*	0.26, 1.5, [0.07–1.0]	0.11, [0.0003–0.415]	0.35, 1.08, [0–1]	0.25, 1.39, [0–1]

[a] CV, coefficient of variation.

[b] For one catchment, parameter *b* reached 4000 (maximum limit in the calibration).

literature (Table 3), while for a small number of cases, the parameters fall outside previously reported ranges.

6.1.1. Parameter *a*: Propensity for Runoff to Occur Before Complete Saturation of the Catchment

[66] Figure 8a shows that values of a < 0.96 are more common for no-snow catchments along the West Coast and southern Appalachians Mountains, and for snow catchments in the Rocky Mountains and Sierra Nevada. No relationship was found between the spatial distribution of *a* and physical descriptors, and only a weak relationship was found with the hydroclimatic variable *E/P* (catchments with *E/P* > 0.48 tend to have *a* > 0.95).

6.1.2. Parameter *b*: Upper Soil Zone Water Holding Capacity

[67] Estimates for *b* vary mainly between 100 mm and 1600 mm and, when scaled by 10 year mean annual catchment precipitation, *b/P* varies between 0.15 and 1.5. For the majority of catchments *b/P* < 0.5, and only 5% have *b/P* > 1.0. For snow dominated catchments in the Rocky Mountains values of *b/P* > 0.70 are common, while values below 0.25 occur for a only few catchments in the Pacific Northwest and lower elevations of the Ohio region (Figure 8b). No relationship was found between the spatial distribution of *b* and physical descriptors, and only weak relationships were found with BF*/P* and BFI (note that BF*/P* = BFI**Q/P*).

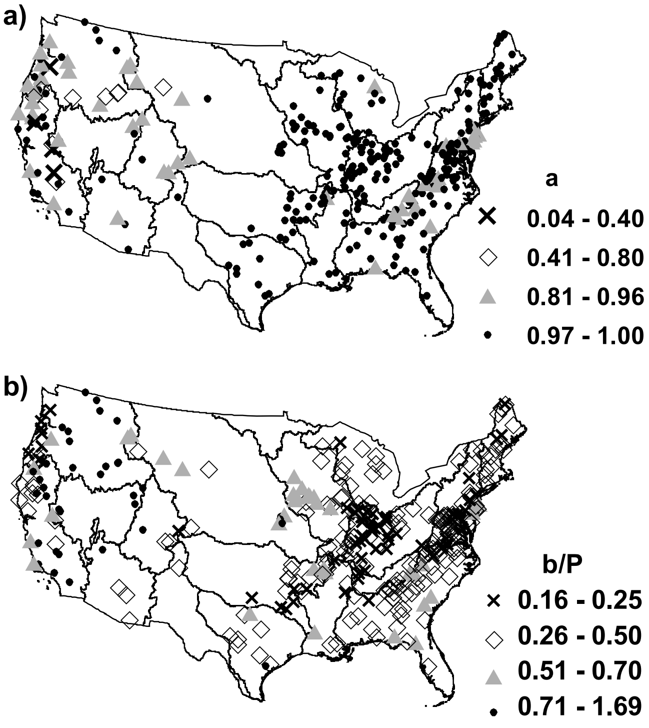

Figure 8. Spatial distribution of parameters *a* and *b*.

6.1.3. Parameters *c* and *d*: Control Degree of Recharge to Groundwater and Its Rate of Release Into the River as Base Flow

[68] These two parameters show varying levels of interaction, depending on activation of the lower zone storage. However, many catchments have calibrated values that fall very close to the extreme points of the feasible ranges [0 1], which correspond to situations which might be better represented using simplified versions of the model. For example *c* ~ 1 corresponds to little or no quick flow contribution to the river, and *d* ~ 0 corresponds to little or no base flow contribution. Figure 9 shows four different model conceptualizations that might be more appropriate for such cases. The standard two–soil tank *abcd* conceptualization (CM-A) remains appropriate when 0.05 < *d* < 0.95; however, this situation occurs for only ~48% of the good catchments. For the remaining catchments, an alternative

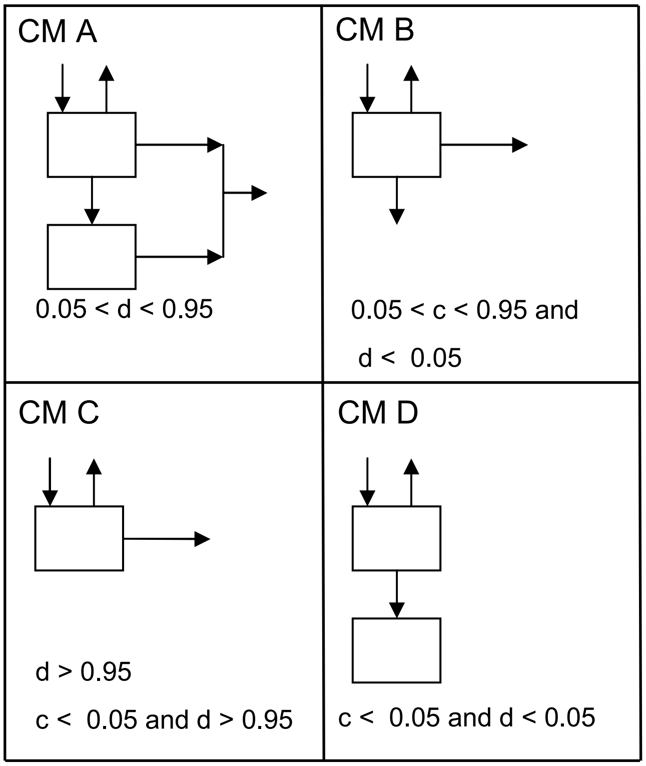

Figure 9. Simplified conceptual models based on the original *abcd* formulation.

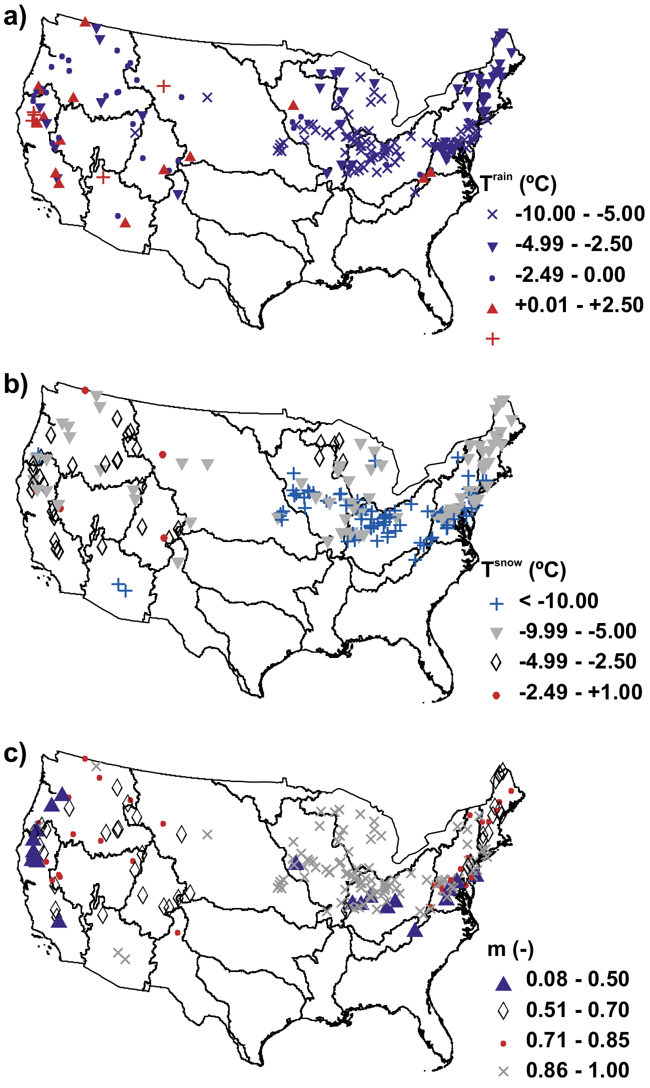

Figure 10. Spatial distribution of the snow parameters (a) T^{rain}, (b) T^{snow}, and (c) m for the catchments classified as good during calibration.

conceptualization may be indicated; for example structure CM-C may be more appropriate for catchments where "very low" values for both c and d were obtained; these correspond to locations where only a single soil zone tank may be required to explain the dynamics of the precipitation-runoff process. In general, *normal* values for these parameters ($0.05 < c$, $d < 0.95$) indicating regions where surface water storage, groundwater storage, and the river are well connected (map not shown), are found mainly along the eastern Appalachians, the lower Upper Mississippi and Ohio, and between the Pacific Northwest and California. This region is relatively humid and energy limited ($\mathrm{PE}/P < 1.2$, $\mathrm{BF}/P > 0.11$ and $Q/P > 0.17$), with more than 34% of total flow comes from groundwater storage ($BFI > 0.34$). Values indicating significant recharge to groundwater, but weak or nonexistent connection between groundwater storage and the river ($d < 0.05$) are predominant in the Great Lakes region and New England. Values indicating weak connection between surface water and groundwater ($c < 0.05$), or where the water table is so high that the distinction between them is not relevant to catchment behavior, are scattered throughout the United States.

6.2. Evaluation of Snow Component Parameters T^{rain}, T^{snow}, and m

[69] As expected, there is strong geographical correspondence between the importance of snow storage and magnitudes of the parameters of the snow component (Figure 10). Values within the expected ranges are obtained for New England (HUC 01) and across the western United States. For Ohio and Upper Mississippi (HUCs 05 and 07), with lower importance of snow storage, the parameters tend to fall close to the boundaries, suggesting problems with the representation of snow processes in the model. In such cases either the snow conceptualization is inappropriate, or the dynamics of snow accumulation and melt are not significant for explaining water balance dynamics at the monthly time scale.

6.2.1. Parameter T^{rain}: Temperature Above Which All Precipitation Falls as Rain

[70] Values between −2.5°C and 0°C are common in the Rocky Mountains, while lower values between −5.0°C and −2.5°C are seen in New England, suggesting a regional bias in temperature for the mountainous catchments since most of the observations are located in the lowlands. Very low T^{rain} (<−7.0°C) results in none of the precipitation being classified as snow. Such values are obtained for most of the catchments in the snow-to-no-snow transition zone in the lower Ohio region, where the snow model hypothesis appears to be ambiguous.

6.2.2. Parameter T^{snow}: Temperature Below Which All Precipitation Falls as Snow

[71] The western United States tends to have higher values of T^{snow} than the East. Unreasonably low values are found for the snow-to-no-snow transition zone.

6.2.3. Parameter m: Snowmelt Rate

[72] Values between 0.5 and 0.8 are found for the western and eastern United States, while values close to 1.0 are found for the snow-to-no-snow transition zone. Beyond this, no strong relationships are found between snow parameters and the physical descriptors examined. While there is a clear positive tendency in the relationship between the temperature thresholds and catchment elevation and slope, and a clear inverse tendency between snowmelt rate and catchment elevation and slope, these relationships are clouded by the interdependence of topography and temperature.

7. Performance Evaluation on Independent Period

[73] We next examine model performance on an independent evaluation period not used for model identification. Specifically, the 764 catchment-calibrated models are evaluated on the decade 1981–1990, using the diagnostic approach to classify catchment model performance. In doing so, we recognize that it is typical for model performance to deteriorate somewhat when going from a calibration period to any period not used for calibration; this happens due to several reasons, including (1) an unavoidable tendency of the MSE approach to force the model to fit some of the noise in the data, (2) a failure of the hypothesis of invariance in catchment behaviors/structures/parameters between periods, (3) deficiencies in the model structure hypothesis (model

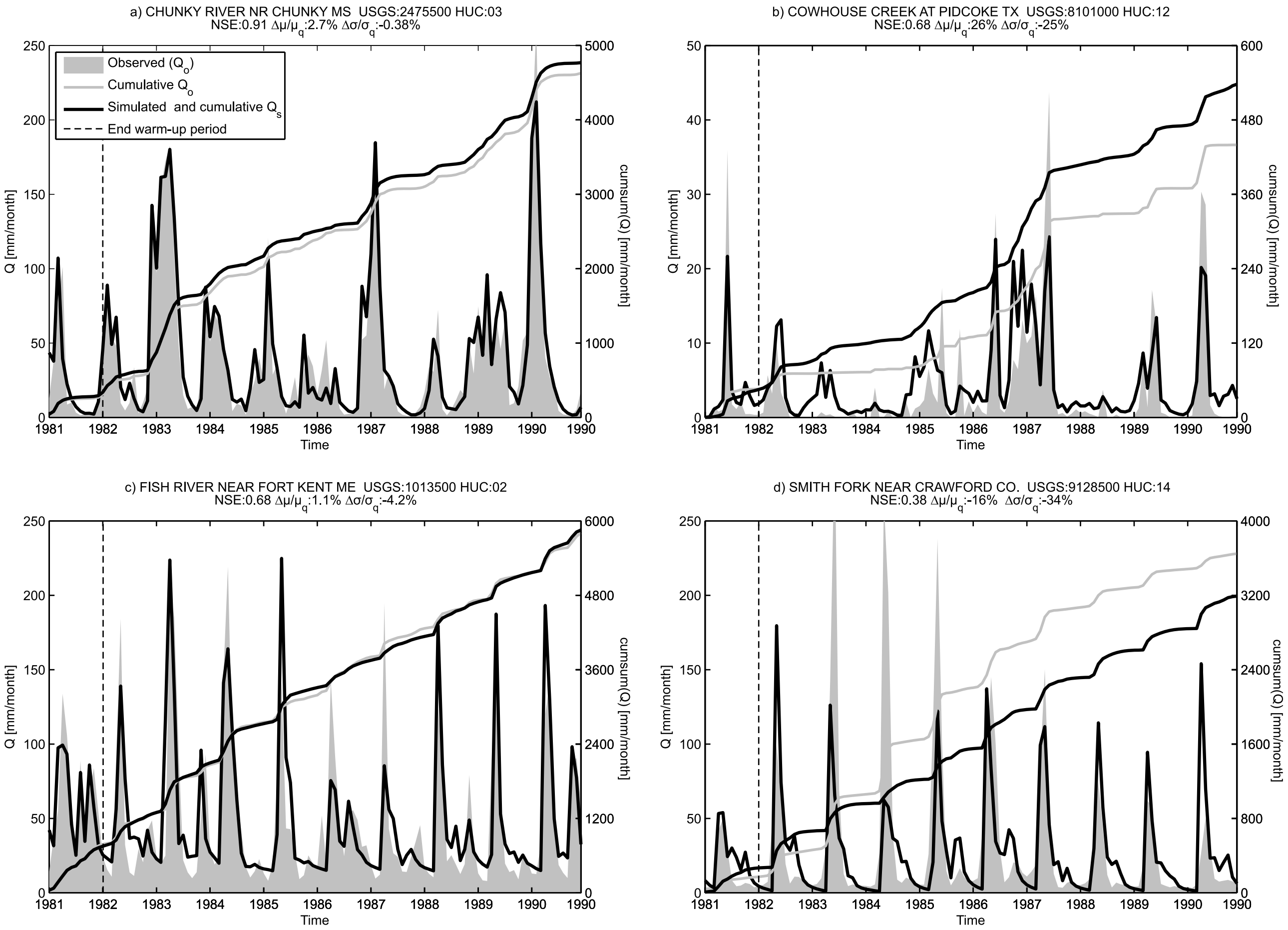

Figure 11. Evaluation period (1981–1990) hydrographs and cumulative hydrographs for four selected catchments: (a) no-snow catchment in the humid southeastern United States having good performance, (b) no-snow catchment in the arid Texas-Gulf region having poor reproduction of water balance and flow variability, (c) snow catchment in the northeastern United States having good reproduction of water balance and flow variability, and (d) snow catchment in the Rocky Mountains having poor reproduction of water balance and flow variability.

structure errors), (4) not properly accounting for uncertainty and random chance in the simulation, identification and evaluation steps, and (5) nonstationarities in the data and its measurement noise properties. Therefore, during calibration we set a higher standard and classify model performance into the two classes good (NSE > 0.75, $|\Delta\mu_q/\mu_q| < 5\%$ and $|\Delta\sigma_q/\sigma_q| < 15\%$) and acceptable/bad (all the rest), and during evaluation we relax the performance standards and classify model performance into the two classes good/acceptable (NSE > 0.67, $|\Delta\mu_q/\mu_q| < 10\%$ and $|\Delta\sigma_q/\sigma_q| < 30\%$) and bad (all the rest).

[74] Further, as an additional test we recalibrate the model for progressively longer time periods (two decades 1951–1970, and three decades 1951–1980) and perform the same examination of evaluation period (1981–1990) performance, to examine the marginal value of increasing the amount of information available during calibration.

[75] Table 4 shows the contingency table of catchment model performance. First we examine the case of no-snow catchments. When we use the first decade of data for model calibration, 53% of the catchments are classified as good during calibration, but only 35% of these also remain good/acceptable during evaluation. As we increase the length of calibration data to two and three decades, the number of catchments classified as good during calibration drops only slightly (from 53%) to 52% and 50% respectively, while the number of these remaining good/acceptable during evaluation increases only slightly (from 35%) to 37% and 40%. Meanwhile the number of catchments that are acceptable/bad during calibration and bad during evaluation remains constant at around 30%. Clearly there is only a small marginal value of using additional data for no-snow catchments if we are trying to gain improvements in model performance. Evaluation period observed and simulated hydrographs for two selected no-snow catchments are presented in Figure 11; the Chunky River (humid southeastern United States) has good performance on all measures, and the Cowhouse Creek (arid Texas-Gulf region) has poor reproduction of water balance and flow variability; catchment locations are indicated in Figure 12.

[76] The evaluation-good/acceptable catchments (Figure 12) are located mainly in the southeast (HUC 3) and the northern Pacific Coast (HUC 18), with ~90% of them in humid regions (having 0.4 < *E/P* < 0.8 and *PE/P* < 1.4), where the *abcd* structure is apparently adequate to represent WB dynamics. The evaluation-bad catchments are located mainly in the

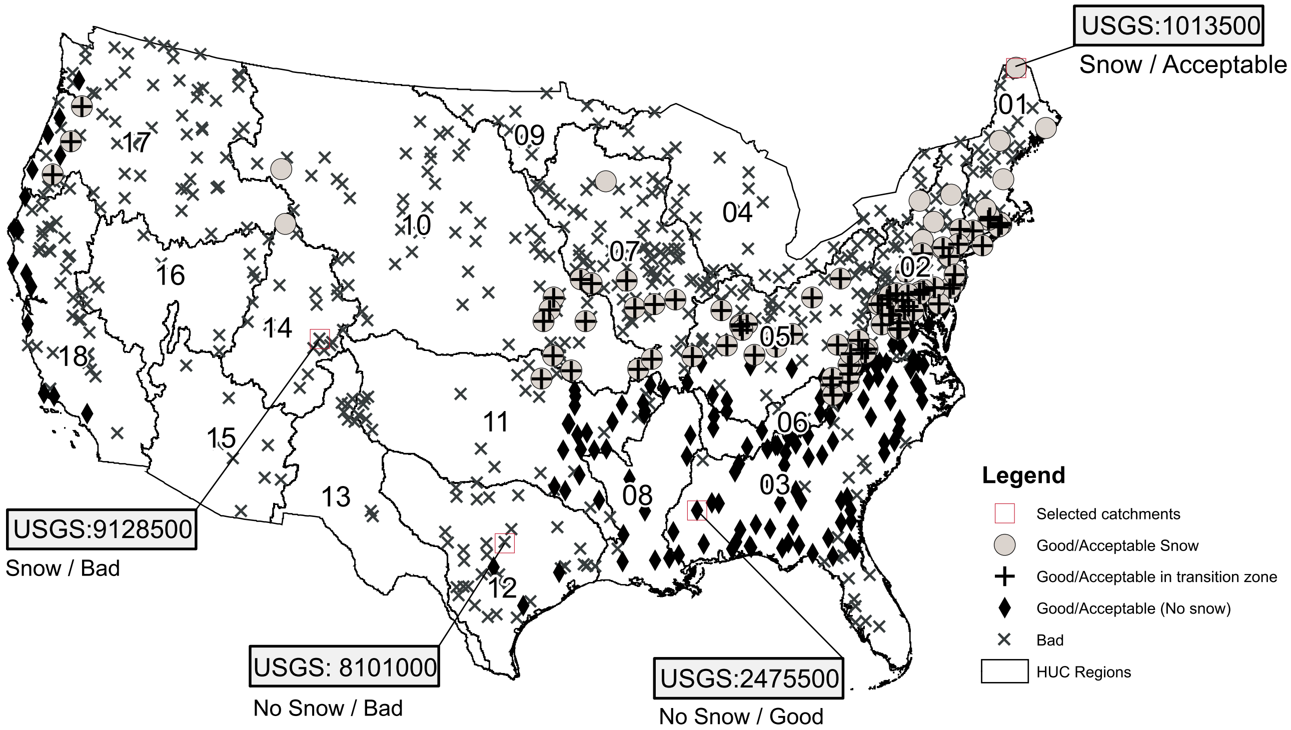

Figure 12. Spatial distribution of the catchments with good or acceptable model performance on the evaluation (1981–1990). Parameters are those from the 30 year calibration period (1951–1980). Selected catchments are marked, for which hydrograph comparisons are shown in Figure 11. USGS 2475500 (no snow) and USGS 1013500 (snow) have good/acceptable performance; USGS 8101000 (no snow) and USGS 9128500 (snow) have bad performance.

Texas-Gulf region (HUC 12) and southern California (HUC 18), with 55% in regions having either high aridity index ($PE/P > 1.4$) or extreme runoff ratios ($Q/P < 0.2$ or $Q/P > 0.6$), where flow intermittency and/or submonthly variability of fluxes are of major importance.

[77] Next we examine the so-called snow catchments (including those in the transition zone where the snow model hypothesis is ambiguous). When the first decade of data is used for model calibration, only 41% of the catchments are “good” during calibration and only 15% of these also remain good/acceptable during evaluation. Meanwhile the number of bad catchments increases from 59% to 85%. As the length of calibration data is increased to two and three decades, the number of catchments classified as good during calibration drops significantly (from 41%) to 29% and 23% respectively, while the number remaining good/acceptable during evaluation increases only slightly (from 15%) to 15% and 18%. Meanwhile the number of catchments that are acceptable/bad during calibration and bad during evaluation increases continuously (from 52%) to 64% and 66%. Clearly the marginal value of additional calibration data is not to gain improvements in model performance, but rather to reveal inadequacies in the model hypothesis and/or inconsistencies in the data.

[78] Overall, it becomes increasingly clear that the model hypothesis for snow catchments is not well supported by the data. Only 18% of the snow catchments show good/acceptable performance during evaluation and most of these (82 of 95) are located in the transition zone (Figure 12) where the representation of snow accumulation and melt is apparently not important to the water balance. Outside of the transition zone, only 13 of the snow catchments show good/acceptable performance during evaluation: ten of these are in New England (HUC 1 and 2), one in the upper Mississippi, and two in the northern Rocky mountains. And contrary to a reviewers suggestion, we find no significant correspondence between catchment size and evaluation period performance in terms of all three criteria: NSE, water balance, and variability; in fact the two largest catchments

Table 4. Contingency Table for Model Performance During Calibration and Evaluation

	Evaluation 1981–1990					
	No Snow			Snow		
	Good/ Acceptable	Bad	Total	Good/ Acceptable	Bad	Total
	Calibration 1951–1960					
Good	35%	17%	53%	8%	32%	41%
Acceptable/Bad	17%	30%	47%	7%	52%	59%
Total	53%	47%	241	15%	85%	523
	Calibration 1951–1970					
Good	37%	15%	52%	8%	21%	29%
Acceptable/Bad	19%	29%	48%	7%	64%	71%
Total	56%	44%	241	15%	85%	523
	Calibration 1951–1970					
Good	40%	10%	50%	7%	16%	23%
Acceptable/Bad	20%	30%	50%	11%	66%	77%
Total	60%	40%	241	18%	82%	523

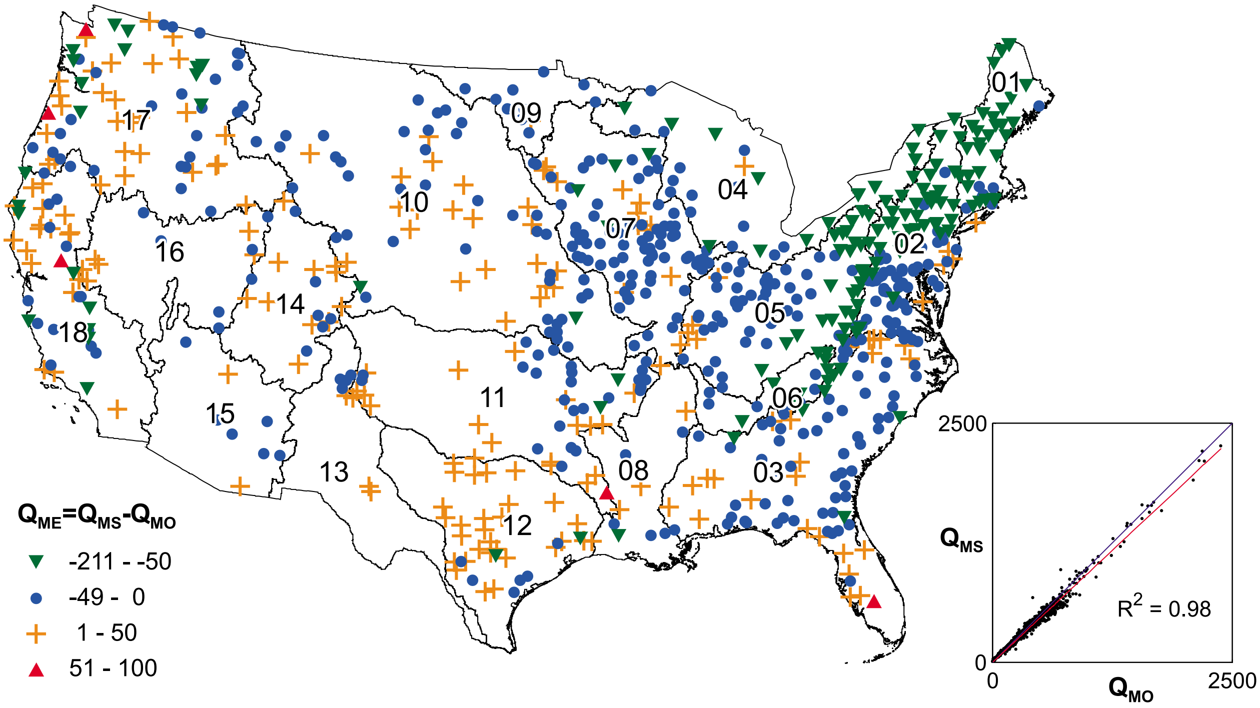

Figure 13. Spatial distribution of the difference between mean annual simulated (Q_{MS}) and observed (Q_{MO}) runoff in mm/yr.

have NSE ~ 0.6, water balance errors <10% and variability errors <10% putting them in the good/acceptable range. Evaluation period observed and simulated hydrographs for two selected snow catchments are presented in Figure 11; the Fish River (northeastern United States) has good reproduction of water balance and flow variability, and the Smith Fork (Rocky Mountains) has poor reproduction of water balance and flow variability; locations are indicated on Figure 12.

[79] Finally, with regard to consistency of model performance across different decades, only 106 of the 339 good first-decade calibration catchments (both snow and no snow) remain good/acceptable when tested in evaluation mode on the three subsequent periods. These are mainly no-snow catchments located along the Pacific Coast and in the southeastern United States. (HUCs 03, 05, 11) excluding Florida. Overall, the analysis suggests the need for an improved approach to monthly time scale water balance modeling across the United States, particularly for arid regions with PE/P > 1.4 and for regions where snow processes contribute significantly to the monthly water balance.

8. Comparison With Previous Studies

[80] Among previous investigations, the most comparable is that of *Hay and McCabe* [2002] who calibrated a similar conceptual model (with different equations) to 44 catchments across the United States. They extrapolated an adjusted R^2 calibration performance to 1646 stations across the United States via multilinear regression based on mean basin elevation and percentage of months with mean basin areal annual runoff <5 mm (R^2 = 0.71); these explanatory variables are related to the catchment descriptors T_a, PE/P and Q/P found to have explanatory power in our study. However, when we apply the same procedure to our data set, we obtain an R^2 of only 0.34 using all 764 catchments, and 0.37 using 39 of the 44 catchments used in their study (the remaining 5 were not in our database). Further, when comparing our adjusted R^2 map to their extrapolated adjusted R^2 map (results not shown), we find generally lower performance in the snow regions and higher performance in the arid regions. This illustrates the dangers of using only a small numbers of catchments to extrapolate results across the United States. However, one important difference between our study and theirs is that *Hay and McCabe* [2002] search for the temperature and precipitation stations giving optimal calibration performance, whereas we assume the PRISM data to adequately correct for topographic effects and other biases. Generally, this comparison leads us to believe that precipitation biases may be significant for many catchments in the HCDN data set, compounded by problems with watershed delineation.

[81] Three other continental scale water balance studies of importance are those of *Milly* [1994], *Arnold et al.* [1999], and *Wolock and McCabe* [1999] who compare 30 year *mean annual* estimates of runoff for the conterminous United States against observations; while all three show scatterplots, only one [*Arnold et al.*, 1999] reports R^2 values, which are on the order of 0.66–0.80. In comparison we obtain a very high R^2 = 0.98 for the 30 year evaluation period (1951–1980), using parameters obtained by calibrating to the 10 year period (1951–1960). Figure 13 shows the spatial distribution of bias in 30 year mean annual runoff estimated by our approach. Catchments in New England and the Appalachian Mountains generally have significant negative bias (ΔQ < −50 mm), while several catchments along the Pacific Coast have significant positive bias (ΔQ > +50 mm). In general our results are consistent with those of the previous studies,

although our biases are generally smaller; in our case 78% of catchments have absolute bias less than 50 mm and 35% have bias less than 10 mm, compared with 45% and 18% reported by *Arnold et al.* [1999]. Note, however, that this method of reporting model bias does not help to reveal significant errors in water balance and flow variability. For example, of the 264 catchments having annual water balance errors smaller than 10 mm, as many as 21% have monthly water balance errors greater than 10% and variability errors greater than 30% causing them to be classified as having bad performance.

9. Summary and Conclusions

[82] The work presented here has explored the challenges involved in estimating water balance models for the conterminous United States at monthly time scale. A unique aspect of our investigation is the use of a large data set of 764 catchments, selected for its comprehensive coverage of hydrogeoclimatic conditions across the United States. Overall our analysis indicates that further iterations of the model development cycle are necessary based on a diagnostic multiple-criteria performance evaluation strategy and a search for the most parsimonious plausible model hypothesis.

[83] In particular, we find that to conduct a robust model evaluation it is not sufficient to rely upon conventional NSE and/or r^2 aggregate statistics of model performance. To have some reasonable degree of confidence that the model can provide hydrologically consistent simulations of catchment behavior, it is also necessary to examine measures of water balance and hydrologic variability (and perhaps others). While this paper examines only the ability to reproduce the first two moments (mean and variance) of the distribution of flows, the mean being related to the overall runoff ratio and the variance being related to the slope of the so-called "flow duration curve," a more robust approach could include other hydrologically relevant variables of interest including indices of temporal dynamics.

[84] Further, it is clear that even if the model displays "good" performance during calibration (in terms of NSE, water balance and variability), this is no guarantee of "acceptable" evaluation period performance. In this regard, our use of longer periods of data for model calibration did not help to improve overall model performance. Therefore, the problem seems not be one of insufficient data length, but rather one of inability of the model structural hypothesis to represent the range of hydrological behaviors experienced across the continental United States.

[85] Overall, we find that the *abcd* model structural hypothesis, implemented in lumped catchment scale fashion at monthly time step, is only suitable for about 35%–40% of the no-snow catchments and 15%–18% of the snow catchments (when augmented with a simple temperature-based parameterization of snow accumulation and melt dynamics), most of the latter being located in the transition zone where snow dynamics appear to be unimportant at the monthly time scale. Model suitability (and hence performance in terms of NSE, water balance and flow variability) has a clear correspondence with geographical characteristics and major hydroclimatic controls; the variables having the greatest power to explain performance being the PE/P ratio (indicating aridity) and mean annual temperature T_a (related to factors including elevation, evapotranspiration, snow processes, etc). In general, the model hypothesis is supported for humid/energy-limited catchments having PE/P < 1.4 located along the West Coast and throughout the East. Meanwhile the hypothesis fails for 30% of no-snow and 65% of snow catchments. These are mainly located in water-limited regions having PE/P > 1.4 located north to south across the dry central United States, the mountainous western United States, and the Northern Plains where the existence of potholes radically impacts hydrological behavior.

[86] For catchments for which the *abcd* model structure appears suitable, the estimated values for parameters *a* and *b* are found to fall comfortably within ranges reported in the literature, with a few exceptions. However, in as many as 50% of these cases the estimated values for parameters *c* and *d* are either not well constrained, or correspond to situations possibly better explained by simplifications to the conceptual model structure. This suggests two things. On the one hand, it seems necessary to find ways to better constrain the values of *c* and *d* during calibration. On the other, a more robust result would likely be obtained by a strategy of testing for appropriate model structural complexity – beginning with only an upper zone water balance reservoir (having only two parameters *a* and *b*) and progressively adding structural complexity as warranted. In this regard, it may be advisable to also test using a multiplier to correct for persistent precipitation biases. Of interest, and somewhat disturbing, is that we find no significant relationships between tested hydroclimatic catchment descriptors and the two parameters *a* and *b* that control the overall water balance behavior (and to some degree the flashiness) of the model response. This issue clearly needs further investigation.

[87] In regard to catchments where the augmented *abcd* model appears suitable, we find strong geographical correspondence between the snow parameters (T^{rain}, T^{snow} and *m*) and the relative importance of the snow storage. Values of the melt rate parameter *m* are found to be between 0.5 and 0.8, which is in contrast to the fixed value of m = 0.5 used for all catchments regardless of location by *McCabe and Wolock* [1999] and *Hay and McCabe* [2002].

[88] For regions where the model hypothesis is found to be unsuitable, a considerable amount of further investigation is necessary to determine the reasons why, and for more suitable structural hypotheses to be proposed and tested. Toward that goal, the results reported in this paper can provide some guidance. Our results clearly suggest that catchments in water-limited regions may need some alternative representation of the mechanisms by which water is partitioned into runoff, evapotranspiration and recharge. Similarly, most regions seem to require more sophisticated representations of the snow accumulation and melt process. It seems quite reasonable to suggest that some representation of the submonthly temporal and/or subcatchment spatial distributions of hydrological fluxes may actually prove to be critical [e.g., *Kling et al.*, 2006], or perhaps some process-based formulation that approximates the terms of the energy balance using daily data [*Walter et al.*, 2005]. And while, in some cases, it may be possible to *parameterize* these subscale processes to enable continued modeling at the monthly catchment scale, in other cases such an approach may fail with the only recourse being to model at some finer spa-

tiotemporal scale. Notwithstanding all this, we should also acknowledge that that several of the catchment descriptors being used in our classification methodology are simply estimates derived via other models, for example, PPS and average SWE, and these have no doubt introduced to the uncertainty into our classification scheme.

10. Recommendations for Future Work

[89] The conventional approach to identification of catchment-scale water balance models needs to be improved. Even when the data are of good quality, the procedure does not suitably constrain the model to give good reproduction of important hydrological behaviors (such as long-term water balance and flow variability) and, as a consequence, reported values of NSE or r^2 can be misleading. Further, we should continue to explore multiple alternative model hypotheses until an appropriate parsimonious structure (and spatiotemporal scale) for each region has been determined. Until these issues are resolved, water balance models of this kind cannot be reliably used to make inferences about the spatial and temporal dynamics of the continental water balance of the United States, or to regionalize model structures and parameters to ungaged locations.

[90] In our ongoing work we seek to refine several elements of the methodology, so as to improve the robustness of the results and establish a more rigorous framework for model development and evaluation. In this regard, a few comments are in order.

10.1. Data Sets

[91] Data sets such as the HCDN constitute excellent test beds for model identification studies, and we must continue to develop their usefulness. Analysis of submonthly and subcatchment scale variability may help to identify locations where the lumped monthly approach is inadequate for representation of catchment scale water balance dynamics. Improved catchment delimitation using more detailed topographic information can help reduce bias errors in precipitation and other fluxes. Similarly, information about gage density could be useful to evaluate problems with the quality of the precipitation input. Consolidation of secondary information can help in development of better catchment descriptors and facilitate evaluation of model performance and parameters. Inclusion of additional indices can facilitate selection of model structures, and may also help to establish theoretical-conceptual limits on acceptable ranges for the hydroclimatic variables being simulated; e.g., it would be useful to know the expected probability distribution of mean annual runoff ratio conditioned on soils, geology, topography, and vegetation information, and the conditions under extreme values might be considered acceptable, and what values might be considered suspicious. Also, for some regions (e.g., the Northern Plains) the model inputs appear to be insufficient, and in high elevation catchments the monthly temperature records may be insufficient for simulating the dynamics of snow. For these regions, and for higher elevations in the western United States, improvements in data acquisition may be needed. And, although not usable in our study since the data is only available starting from 1999, remotely sensed MODIS based snow cover mapping products could be useful in determining the extent to which a basin is impacted by snow.

10.2. Catchment Classification

[92] Our results reinforce the notion that a simple classification system based on common hydroclimatological descriptors (such as the ones used in the climate classification systems) can help in initial categorization of catchments and selection of appropriate conceptual model structures and parameterizations. Different levels of model performance within each classification could then be related to catchment characteristics such as the predominant soil type, geology and landscape. The goal will be a more sophisticated understanding of the methods by which model structures, and parameter values, can be regionalized. In this regard, the classification of snow regions needs considerable attention. The diagnostic process can benefit from more sophisticated methods for evaluating effects of snow storage on catchment dynamics, and could include the use of snow information from reanalysis data sets. Such a classification could be used to extend the HCDN descriptors to help in identifying the relative importance of snow storage within a catchment.

10.3. Model Complexity and Parameter Uncertainty

[93] There is a need for the model identification process to diagnostically consider alternative model structures with varying degrees of complexity, as suggested decades ago by *Nash and Sutcliffe* [1970]. For example, it might be useful to follow a progressive approach to introducing model complexity, as appropriate to the particular hydroclimatology and data conditions of the catchment region. Specific enhancements can include precipitation multipliers to correct for biases in the precipitation, the addition of rudimentary sublimation schemes for the snow storage, and accounting for recharge loss to groundwater. Further enhancements can include the inclusion of additional storages to account for soil water redistribution and/or routing, as necessary. For cases where large model errors tend to persist, machine learning and data mining tools could be used to explore the reasons. In this regard, and to help in correlation of parameters with catchment and hydroclimatic descriptors, it would be useful to quantify the precision of parameter estimates. Estimates of the posterior parameter covariance matrix could be obtained in several ways, including the classical approach of explicit or finite difference computation of derivatives [*Gupta and Sorooshian*, 1985], use of Monte Carlo Markov chain based optimization methods [*Vrugt et al.*, 2003], or by application of bootstrapping in the computation of the performance criteria [*Ebtehaj et al.*, 2010].

10.4. Metrics for Assessing Model Performance

[94] Future work should include an appropriate set of robust metrics that specifically target relevant hydroclimatological behaviors to be reproduced by the model [*Gupta et al.*, 2008, 2009]. Further, the value of a model in any region, should be evaluated against an appropriate baseline [*Schaefli and Gupta*, 2007], evaluation against the long-term mean, as done by NSE, does not constitute a powerful test of model performance. Such baselines can include extensions of hydroclimatological frameworks [*Milly*, 1994] to generate metrics that progressively test with regards to the main patterns of seasonal variation. Similarly, a more robust test of model performance should take into account model complexity [e.g., *Ye et al.*, 2008].

Table A1. Explanatory Predictors and Decision Trees Used to Explore the Results

Variable	PE/P	Q/P	E/P	Q_{99}	$T_{\text{Win. Max}}$	Q	P	(P-Q)/P	E_a	E May	Solar	T_a	PPS	SWE	T_{range}	$T_{\max}$	$T_{\text{Jan. Min}}$	$T_{\text{Ann. Min}}$	CDD	HDD	Snow	Elevation	Area	Slope	Permeability	Slope	SWCH	Aspect	MEAN BFI	BF/P
Decision Tree																														
DT01	X	X		X	X							X	X																	
DT02	X	X		X	X			X					X									X	X	X						
DT03	X	X	X			X	X	X	X	X																				
DT04	X	X		X				X					X										X						X	X
DT05					X								X	X	X	X	X	X	X	X	X	X		X						
DT06	X							X				X												X					X	X
DT07		X						X											X	X		X	X							
DT08																									X	X	X			
DT09																						X	X	X						
DT10	X				X			X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X
Data Source																														
Time Series 40Y.	X	X	X	X	X																									
Time Series Eval						X	X	X																						
HCDN Database									X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

10.5. Calibration Strategy

[95] A calibration strategy based simply on optimizing a mean squared error type criterion appears to be inadequate for water balance modeling at the monthly time scale. Even in this relatively simple situation, a more sophisticated strategy based on multiple-criteria analysis appears to be necessary to properly reproduce the different aspects of catchment behavior to be simulated by the model. Further, an important test of model robustness is to ensure that the “optimal” structure and parameters do not change with time, or do so in a well understood fashion that can be predicted using available indicators.

[96] Finally, with this paper, we hope to encourage more rigorous evaluation and diagnosis of hydrologic models, using the large catchment data sets available in the United States to evaluate and discuss model hypotheses and identification methods, and to identify locations where existing model representations are not adequate. Our data infrastructure is developing rapidly [*Beran and Piasecki*, 2008] and robust methods to exploit the rapidly growing sources and quantities of information need to be developed. The days where a modeling study is based on analysis of the data from a “single” catchment are (or should be) rapidly drawing to a close, and methods for calibrating and analyzing the results from hundreds (even thousands) of catchments under different conditions need to become part of our common practice. While analytical tools to mine the data are now available, we need to continue formalizing our ability to consolidate hydrological knowledge and to identify gaps in understanding. Tools for handling large numbers of catchments can be improved by developing automated algorithms for defining clusters of stations having similar input-output mapping structures, testing different benchmarks to evaluate model assumptions, and for evaluating additional catchment signatures and indices. As always, we invite dialog on these and related issues of model identification.

Appendix A: Decision Trees Used for the Analysis of Results

[97] Classification and regression trees (CART) are a data mining technique used to explore nonparametric relations within a data set [*Kumar et al.*, 2006]. CART algorithms in MATLAB (R2008b) were used to investigate the strength of relationships between watershed descriptors and error components. Explanatory predictors from each catchment (columns in Table A1) were tabulated along with corresponding classes, numerical values of error components and model

Table A2. Classes and Regression Values Investigated With the Trees

Type	Classes/Regression
NSE	Good, acceptable or bad; NSE regression
Performance classification	Good, acceptable or bad; good or not good
Flow bias	Good+, acceptable+, bad+, good-bias-, acceptable- or bad-; $\%\Delta\mu_q/\mu_q$ regression value
Flow variability	Good, acceptable or bad; $\%\Delta\sigma_q/\sigma_q$ regression
Parameters	Regressions for a, b, c, and d; CM-A,CM-B, CM-C or CM-D
Snow parameters	Regression for T^{rain}, m, or T^{snow}

Table A3. Filters Used to Select the Catchments to Construct the Trees

Domains	Filters
General	All stations
Predominant snow dynamics	No snow stations
	Snow stations
	Transition zones
Spatial	HUCs (level 1)
	Groups of HUC's (west, central, east)
Performance classification	Good stations
	Good and bad stations
	Acceptable stations
	Bad stations
Error components	Acceptable or poor flow bias
	Acceptable or poor flow variability

parameters (Table A2). Decision trees were constructed by examining different combinations of climatic and physical predictors (rows in Table A1) to search for those having discriminatory power. At first, the trees were used to analyze relationships between model performance and existing HCDN catchment descriptors [*Vogel and Sankarasubramanian*, 2005]. Later, the analysis was augmented with descriptors computed directly from the time series data, and the hydrologic landscape regions classification [*Wolock et al.*, 2004].

[98] To obtain the optimal node level for each tree, and to avoid over fitting, errors from tree resubstitution and tenfold cross validation were compared. Once relevant variables were identified, visual inspection of scatter and parallel plots was used to check the results.

[99] To facilitate the analysis, and to allow for variation in catchment characteristics and processes, the analysis was performed separately for catchments in different classes/domains (Table A3). For example, predictors of performance were explored independently for different snow regions of the United States. Similarly when exploring relationships between model parameters and descriptors, only locations with good performance were used.

[100] **Acknowledgments.** We wish to express our sincere gratitude to the Editor, Associate Editor, and several reviewers who provided very valuable comments and suggestions that resulted in significant improvements to the manuscript. This work was supported in part by the Hydrology Laboratory of the National Weather Service (grant NA04NWS4620012), by the Center for Sustainability of semi-Arid Hydrology and Riparian Areas (SAHRA) under the STC program of the National Science Foundation (agreement EAR 9876800), by the National Aeronautical and Space Administration under grant NNA06CN09A, and by the Salt River Project.

References

Acreman, M. C., and C. D. Sinclair (1986), Classification of drainage basins according to their physical characteristics; an application for flood frequency analysis in Scotland, *J. Hydrol.*, *84*, 365–380, doi:10.1016/0022-1694(86)90134-4.

Alley, W. M. (1984), Treatment of evapotranspiration, soil moisture accounting, and aquifer recharge in monthly water balance models, *Water Resour. Res.*, *20*, 1137–1149, doi:10.1029/WR020i008p01137.

Alley, W. M. (1985), Water balance models in one-month-ahead streamflow forecasting, *Water Resour. Res.*, *21*, 597–606, doi:10.1029/WR021i004p00597.

Andreassian, V., J. Lerat, C. Loumagne, T. Mathevet, C. Michel, L. Oudin, and C. Perrin (2007), What is really undermining hydrologic science today?, *Hydrol. Processes*, *21*, 2819–2822, doi:10.1002/hyp.6854.

Arnell, N. W. (1999), Climate change and global water resources, *Global Environ. Change*, *9*, S31–S49, doi:10.1016/S0959-3780(99)00017-5.

Arnold, J. G., R. Srinivasan, R. S. Muttiah, and P. M. Allen (1999), Continental scale simulation of the hydrologic balance, *J. Am. Water Resour. Assoc.*, *35*, 1037–1051, doi:10.1111/j.1752-1688.1999.tb04192.x.

Bai, Y., T. Wagener, and P. Reed (2009), A top-down framework for watershed model evaluation and selection under uncertainty, *Environ. Modell. Software*, *24*, 901–916, doi:10.1016/j.envsoft.2008.12.012.

Beran, B., and M. Piasecki (2008), Availability and coverage of hydrologic data in the US geological survey National Water Information System (NWIS) and US Environmental Protection Agency Storage and Retrieval System (STORET), *Earth Sci. Inf.*, *1*, 119–129, doi:10.1007/s12145-008-0015-2.

Berger, K. P., and D. Entekhabi (2001), Basin hydrologic response relations to distributed physiographic descriptors and climate, *J. Hydrol.*, *247*, 169–182, doi:10.1016/S0022-1694(01)00383-3.

Blöschl, G. (2006), Hydrologic synthesis: Across processes, places, and scales, *Water Resour. Res.*, *42*, W03S02, doi:10.1029/2005WR004319.

Budyko, M. I. (1974), *Climate and Life*, Academic, San Diego, Calif.

Clark, M. P., A. G. Slater, D. E. Rupp, R. Woods, J. A. Vrugt, H. V. Gupta, T. Wagener, and L. E. Hay (2008), Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models, *Water Resour. Res.*, *44*, W00B02, doi:10.1029/2007WR006735.

Daly, C., R. P. Neilson, and D. L. Phillips (1994), A statistical-topographic model for mapping climatological precipitation over mountainous terrain, *J. Appl. Meteorol.*, *33*, 140–158, doi:10.1175/1520-0450(1994)033<0140:ASTMFM>2.0.CO;2.

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

Duan, Q., J. Schaake, V. Andreassian, S. Franks, G. Goteti, H. V. Gupta, Y. M. Gusev, F. Habets, A. Hall, and L. Hay (2006), Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops, *J. Hydrol.*, *320*, 3–17, doi:10.1016/j.jhydrol.2005.07.031.

Dunne, T. (1983), Relation of field studies and modeling in the prediction of storm runoff, *J. Hydrol.*, *65*, 25–48, doi:10.1016/0022-1694(83)90209-3.

Ebtehaj, M., H. Moradkhani, and H. V. Gupta (2010), Improving robustness of hydrologic parameter estimation by the use of moving block bootstrap, *Water Resour. Res.*, doi:10.1029/2009WR007981, in press.

Euliss, N. H., D. A. Wrubleski, and D. M. Mushet (1999), Wetlands of the Prairie Pothole region: Invertebrate species composition, ecology, and management, in *Invertebrates in Freshwater Wetlands of North America: Ecology and Management*, edited by D. P. Batzer, R. B. Rader, and S. A. Wissinger, pp. 471–514, John Wiley, New York.

Fernandez, W., R. M. Vogel, and A. Sankarasubramanian (2000), Regional calibration of a watershed model, *Hydrol. Sci. J.*, *45*, 689–707, doi:10.1080/02626660009492371.

Gleick, P. H. (1987), The development and testing of a water balance model for climate impact assessment: Modeling the Sacramento Basin, *Water Resour. Res.*, *23*, 1049–1061, doi:10.1029/WR023i006p01049.

Guo, S., J. Wang, L. Xiong, A. Ying, and D. Li (2002), A macro-scale and semi-distributed monthly water balance model to predict climate change impacts in China, *J. Hydrol.*, *268*, 1–15, doi:10.1016/S0022-1694(02)00075-6.

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

Gupta, H. V., H. Kling, K. K. Yilmaz, and G. F. Martinez (2009), Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, *J. Hydrol.*, *377*, 80–91, doi:10.1016/j.jhydrol.2009.08.003.

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

Hargreaves, G. H., and Z. A. Samani (1982), Estimating potential evapotranspiration, *J. Irrig. Drain. Eng.*, *108*, 225–230.

Hay, L. E., and G. J. McCabe (2002), Spatial variability in water-balance model performance in the conterminous United States, *J. Am. Water Resour. Assoc.*, *38*, 847–860, doi:10.1111/j.1752-1688.2002.tb01001.x.

Hughes, M. G., and D. A. Robinson (1996), Historical snow cover variability in the Great Plains region of the USA: 1910 through to 1993, *Int. J. Climatol.*, *16*, 1005–1018, doi:10.1002/(SICI)1097-0088(199609)16:9<1005::AID-JOC63>3.0.CO;2-0.

Institute of Hydrology (Ed.) (1980), *Low Flow Studies*, Inst. of Hydrol., Wallingford, U. K.

Kling, H., and H. P. Nachtnebel (2009), A method for the regional estimation of runoff separation parameters for hydrological modelling, *J. Hydrol.*, *364*, 163–174, doi:10.1016/j.jhydrol.2008.10.015.

Kling, H., J. Fürst, and H. P. Nachtnebel (2006), Seasonal, spatially distributed modelling of accumulation and melting of snow for computing runoff in a long-term, large-basin water balance model, *Hydrol. Processes*, *20*, 2141–2156, doi:10.1002/hyp.6203.

Kumar, P., M. Folk, J. C. Alameda, M. Markus, and P. Bajcsy (2006), *Hydroinformatics: Data Integrative Approaches in Computation, Analysis, and Modeling*, CRC Press, Boca Raton, Fla.

Lins, H. F. (1997), Regional streamflow regimes and hydroclimatology of the United States, *Water Resour. Res.*, *33*, 1655–1667, doi:10.1029/97WR00615.

Makhlouf, Z., and C. Michel (1994), A two-parameter monthly water balance model for French watersheds, *J. Hydrol.*, *162*, 299–318, doi:10.1016/0022-1694(94)90233-X.

Martinez, G. F. (2007), Diagnostic evaluation of watershed models, M.S. thesis, Univ. of Ariz., Tucson.

Maurer, E. P., A. W. Wood, J. C. Adam, D. P. Lettenmaier, and B. Nijssen (2002), A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States, *J. Clim.*, *15*, 3237–3251, doi:10.1175/1520-0442(2002)015<3237:ALTHBD>2.0.CO;2.

McCabe, G. J., and D. M. Wolock (1999), General-circulation-model simulations of future snowpack in the western United States, *J. Am. Water Resour. Assoc.*, *35*, 1473–1484, doi:10.1111/j.1752-1688.1999.tb04231.x.

McDonnell, J. J., and R. Woods (2004), On the need for catchment classification, *J. Hydrol.*, *299*, 2–3.

Milly, P. C. D. (1994), Climate, soil water storage, and the average annual water balance, *Water Resour. Res.*, *30*, 2143–2156, doi:10.1029/94WR00586.

Milly, P. C. D., J. Betancourt, M. Falkenmark, R. M. Hirsch, Z. W. Kundzewicz, D. P. Lettenmaier, and R. J. Stouffer (2008), Stationarity is dead: Whither water management?, *Science*, *319*, 573–574, doi:10.1126/science.1151915.

Mouelhi, S., C. Michel, C. Perrin, and V. Andréassian (2006), Stepwise development of a two-parameter monthly water balance model, *J. Hydrol.*, *318*, 200–214, doi:10.1016/j.jhydrol.2005.06.014.

Nash, J. E., and J. V. Sutcliffe (1970), River flow forecasting through conceptual models part I—A discussion of principles, *J. Hydrol.*, *10*, 282, doi:10.1016/0022-1694(70)90255-6.

National Climatic Data Center (2002), *Climate Atlas of the United States* [CD-ROM], Natl. Oceanic and Atmos. Admin., U.S. Dep. of Commer., Asheville, N. C.

Neuman, S. P. (2003), Maximum likelihood Bayesian averaging of uncertain model predictions, *Stochastic Environ. Res. Risk Assess.*, *17*, 291–305, doi:10.1007/s00477-003-0151-7.

Perrin, C., C. Michel, and V. Andreassian (2001), Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments, *J. Hydrol.*, *242*, 275–301, doi:10.1016/S0022-1694(00)00393-0.

Perrin, C., V. Andréassian, C. R. Serna, T. Mathevet, and N. Le Moine (2008), Discrete parameterization of hydrological models: Evaluating the use of parameter sets libraries over 900 catchments, *Water Resour. Res.*, *44*, W08447, doi:10.1029/2007WR006579.

Rodell, M., P. R. Houser, U. Jambor, J. Gottschalck, K. Mitchell, C. J. Meng, K. Arsenault, B. Cosgrove, J. Radakovich, and M. Bosilovich (2004), The global land data assimilation system, *Bull. Am. Meteorol. Soc.*, *85*, 381–394, doi:10.1175/BAMS-85-3-381.

Sankarasubramanian, A., and R. M. Vogel (2002a), Annual hydroclimatology of the United States, *Water Resour. Res.*, *38*(6), 1083, doi:10.1029/2001WR000619.

Sankarasubramanian, A., and R. M. Vogel (2002b), Comment on the paper: "Basin hydrologic response relations to distributed physiographic descriptors and climate" by Karen Plaut Berger, Dara Entekhabi, 2001. Journal of Hydrology 247, 169–182, *J. Hydrol.*, *263*, 257–261.

Savenije, H. H. G. (2008), The art of hydrology, *Hydrol. Earth Syst. Sci. Discuss.*, *5*, 3157–3167, doi:10.5194/hessd-5-3157-2008.

Schaefli, B., and H. V. Gupta (2007), Do Nash values have value?, *Hydrol. Processes*, *21*, 2075–2080, doi:10.1002/hyp.6825.

Seaber, P. R., F. P. Kapinos, and G. L. Knapp (1987), Hydrologic unit maps, *U.S. Geol. Surv. Water Supply Pap., 2294*, 61 pp.

Sivapalan, M. (2005), Pattern, process and function: Elements of a unified theory of hydrology at the catchment scale, in *Encyclopedia of Hydrological Sciences*, vol. 13, edited by M. G. Anderson, pp. 193–219, John Wiley, Chichester, U. K.

Sivapalan, M., J. Schaake, and J. Sapporo (2003), PUB science and implementation plan, Int. Assoc. of Hydrol. Sci., Gentbrugge, Belgium.

Slack, J. R., A. M. Lumb, and J. M. Landwehr (1993), Hydro-Climatic Data Network (HCDN) streamflow data set, 1874–1988, *U.S. Geol. Surv. Water Resour. Invest. Rep., 93-4076*.

Smith, M. B., D. J. Seo, V. I. Koren, S. M. Reed, Z. Zhang, Q. Duan, F. Moreda, and S. Cong (2004), The distributed model intercomparison project (DMIP): Motivation and experiment design, *J. Hydrol.*, *298*, 4–26, doi:10.1016/j.jhydrol.2004.03.040.

Sturm, M., J. Holmgren, and G. E. Liston (1995), A seasonal snow cover classification system for local to global applications, *J. Clim.*, *8*, 1261–1283, doi:10.1175/1520-0442(1995)008<1261:ASSCCS>2.0.CO;2.

Thomas, H. A. (1981), Improved methods for national water assessment: Final report, *U.S. Geol. Surv. Water Resour. Contract WR15249270*, 44 pp.

Thomson, A. M., N. J. Rosenberg, R. C. Izaurralde, and R. A. Brown (2005), Climate change impacts for the conterminous USA: An integrated assessment—Part 2: Models and validation, *Clim. Change*, *69*, 27–41, doi:10.1007/s10584-005-3609-4.

Thornthwaite, C. W. (1948), An approach toward a rational classification of climate, *Geogr. Rev.*, *38*, 55–94, doi:10.2307/210739.

Vandewiele, G. L., and A. Elias (1995), Monthly water balance of ungauged catchments obtained by geographical regionalization, *J. Hydrol.*, *170*, 277–291, doi:10.1016/0022-1694(95)02681-E.

Vandewiele, G. L., C.-Y. Xu, and N.-L. Win (1992), Methodology and comparative study of monthly water balance models in Belgium, China and Burma, *J. Hydrol.*, *134*, 315–347, doi:10.1016/0022-1694(92)90041-S.

Vogel, R. M., and A. Sankarasubramanian (2005), USGS Hydro-Climatic Data Network (HCDN): Monthly Climate Database, 1951–1990, vol. 2005, http://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=810, Oak Ridge Natl. Lab. Distributed Active Arch. Cent., Oak Ridge, Tenn.

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

Wagener, T., M. Sivapalan, P. A. Troch, and R. Woods (2007), Catchment classification and hydrology similarity, *Geogr. Compass*, *1*(4), 901–931, doi:10.1111/j.1749-8198.2007.00039.x.

Walter, M. T., E. S. Brooks, D. K. McCool, L. G. King, M. Molnau, and J. Boll (2005), Process-based snowmelt modeling: Does it require more input data than temperature-index modeling?, *J. Hydrol.*, *300*, 65–75, doi:10.1016/j.jhydrol.2004.05.002.

Wang, G., J. Xia, and J. Chen (2009), Quantification of effects of climate variations and human activities on runoff by a monthly water balance model: A case study of the Chaobai River basin in northern China, *Water Resour. Res.*, *43*, W00A11, doi:10.1029/2007WR006768.

Wardrop, D. H., J. A. Bishop, M. Easterling, K. Hychka, W. Myers, G. P. Patil, and C. Taillie (2005), Use of landscape and land use parameters for classification and characterization of watersheds in the mid-Atlantic across five physiographic provinces, *Environ. Ecol. Stat.*, *12*, 209–223, doi:10.1007/s10651-005-1042-5.

Winter, T. C. (2001), The concept of hydrologic landscapes, *J. Am. Water Resour. Assoc.*, *37*, 335–349, doi:10.1111/j.1752-1688.2001.tb00973.x.

Wolock, D. M., and G. J. McCabe (1999), Explaining spatial variability in mean annual runoff in the conterminous United States, *Clim. Res.*, *11*, 149–159, doi:10.3354/cr011149.

Wolock, D. M., T. C. Winter, and G. McMahon (2004), Delineation and evaluation of hydrologic-landscape regions in the United States using geographic information system tools and multivariate statistical analyses, *Environ. Manage. N. Y.*, *34*, *Suppl. 1*, S71–S88, doi:10.1007/s00267-003-5077-9.

Xiong, L., and S. Guo (1999), A two-parameter monthly water balance model and its application, *J. Hydrol.*, *216*, 111–123, doi:10.1016/S0022-1694(98)00297-2.

Xu, C.-Y., and V. P. Singh (1998), A review on monthly water balance models for water resources investigations, *Water Resour. Manage.*, *12*, 20–50, doi:10.1023/A:1007916816469.

Xu, C.-Y., and G. L. Vandewiele (1995), Parsimonious monthly rainfall-runoff models for humid basins with different input requirements, *Adv. Water Resour.*, *18*, 39–48, doi:10.1016/0309-1708(94)00017-Y.

Yadav, M., T. Wagener, and H. V. Gupta (2007), Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins, *Adv. Water Resour.*, *30*, 1756–1774, doi:10.1016/j.advwatres.2007.01.005.

Ye, M., P. D. Meyer, and S. P. Neuman (2008), On model selection criteria in multimodel analysis, *Water Resour. Res.*, *44*, W03428, doi:10.1029/2008WR006803.

Zhang, L., N. Potter, K. Hickel, Y. Zhang, and Q. Shao (2008), Water balance modeling over variable time scales based on the Budyko framework—Model development and testing, *J. Hydrol.*, *360*, 117–131, doi:10.1016/j.jhydrol.2008.07.021.

H. V. Gupta and G. F. Martinez, Department of Hydrology and Water Resources, University of Arizona, Tucson, AZ 85721, USA. (gfmb@hwr.arizona.edu)