When are multiobjective calibration trade-offs in hydrologic models meaningful?

J. B. Kollat,[1] P. M. Reed,[1] and T. Wagener[1]

Received 19 October 2011; revised 30 January 2012; accepted 8 February 2012; published 21 March 2012.

[1] This paper applies a four-objective calibration strategy focusing on peak flows, low flows, water balance, and flashiness to 392 model parameter estimation experiment (MOPEX) watersheds across the United States. Our analysis explores the influence of model structure by analyzing how the multiobjective calibration trade-offs for two conceptual hydrologic models, the Hydrology Model (HYMOD) and the Hydrologiska Byråns Vattenbalansavdelning (HBV) model, compare for each of the 392 catchments. Our results demonstrate that for modern multiobjective calibration frameworks to identify any meaningful measure of model structural failure, users must be able to carefully control the precision by which they evaluate their trade-offs. Our study demonstrates that the concept of epsilon-dominance provides an effective means of attaining bounded and meaningful hydrologic model calibration trade-offs. When analyzed at an appropriate precision, we found that meaningful multiobjective trade-offs are far less frequent than prior literature has suggested. However, when trade-offs do exist at a meaningful precision, they have significant value for supporting hydrologic model selection, distinguishing core model deficiencies, and identifying hydroclimatic regions where hydrologic model prediction is highly challenging.

Citation: Kollat, J. B., P. M. Reed, and T. Wagener (2012), When are multiobjective calibration trade-offs in hydrologic models meaningful?, *Water Resour. Res.*, *48*, W03520, doi:10.1029/2011WR011534.

1. Introduction

[2] *Gupta et al.* [1998] carefully frame how multiobjective hydrologic calibration can be used as a means to extract more information from observed time series by using several measures of performance. This view has evolved to the more broadly defined area of model diagnostics [*Gupta et al.*, 2008] where evaluation and identification of parameters are conditional on multiple hydrologic responses (e.g., high flow, low flow, water balance, flashiness, etc.). Beyond hydrologic modeling, the mathematical challenges and multiobjective nature of model identification has long been recognized in the water resources literature [*Emsellem and de Marsily*, 1971; *Neuman*, 1973; *Sun*, 1994]. The core assumption in multiobjective hydrologic model calibration is that limitations in model structure, in combination with the parameterization conflicts, will yield trade-offs across the multiple error measures used to represent a suite of hydrologic responses. For example, many hydrologic modeling studies have often focused on the trade-offs between high and low streamflow responses [*Bekele and Nicklow*, 2007; *Boyle et al.*, 2000; *Cheng et al.*, 2002; *Fenicia et al.*, 2007; *Gill et al.*, 2006; *Khu and Madsen*, 2005; *Khu et al.*, 2008; *Madsen*, 2000; *Madsen et al.*, 2002; *Tang et al.*, 2006, 2007; *van Griensven and Bauwens*, 2003; *Vrugt et al.*, 2003; *Wagener et al.*, 2001; *Yapo et al.*, 1998]. The mathematical principle of multiobjective optimality (or Pareto optimality) is fundamentally based on conflict. For example, Pareto optimal hydrologic model calibration trade-offs only exist if a parameter sets performance with respect to high flows cannot be improved without degrading its performance relative to low flows. If this conflict does not exist, then a single parameterization exists that optimizes both high-flow response and low-flow response, causing the two-objective trade-off to collapse to this single point. In general, an M-objective calibration problem will have at maximum an $(M-1)$ dimensional Pareto front assuming parameterization conflicts exist for all combinations of the objectives [*Das*, 1999; *di Pierro et al.*, 2007; *Khu and Madsen*, 2005; *Teytaud*, 2007]. If, however, no parameterization conflicts exist, then the M-objective multiobjective formulation's solution will collapse to a zero dimensional geometry (i.e., a single parameter set that optimizes all objectives).

[3] Starting with the early multiobjective calibration work [*Gupta et al.*, 1998; *Yapo et al.*, 1998] and the large body of literature it has inspired over the past decade (for a comprehensive review see *Efstratiadis and Koutsoyiannis* [2010]), the existence and meaning of hydrologic calibration trade-offs have largely been discussed as representing structural deficiencies in conceptual hydrologic models. Detailed discussions of these issues are present in the early studies [e.g., see *Gupta et al.*, 1998; *Seibert*, 2000; *Seibert and McDonnell*, 2002; *Wagener et al.*, 2003]. Ideally, multiobjective trade-offs should not exist for hydrologic models. Their presence represents a failure to identify a single set of parameters that allow the model to optimize simultaneously all of the specified measures of performance. For

[1] Department of Civil and Environmental Engineering, Pennsylvania State University, University Park, Pennsylvania, USA.

Copyright 2012 by the American Geophysical Union
0043-1397/12/2011WR011534

conceptual hydrologic models, the presence or absence of calibration trade-offs is strongly shaped by the ability of the model to capture the suite of user specified objectives across diverse watersheds [*van Werkhoven et al.*, 2008, 2009], the complexity of each of the component objectives' response surfaces (i.e., multimodality, see *Duan et al.* [1992], *Kavetski and Kuczera* [2007], *Kavetski et al.* [2006], and *Tang et al.* [2006]), and the numerical precision at which output errors are evaluated [*Kavetski and Clark*, 2010, 2011].

[4] Given the uncertainties and observation errors implicit to hydrologic modeling, our study demonstrates that prior multiobjective calibration exercises have suffered from excessively precise trade-off analysis [*Boyle et al.*, 2000; *Gupta et al.*, 1998; *Tang et al.*, 2006, 2007; *van Werkhoven et al.*, 2009; *Yapo et al.*, 1998], providing a false sense of calibration trade-offs. The principle of nondomination sorting underlies multiobjective optimization, particularly for approaches based on multiobjective evolutionary algorithms (MOEAs) (see the review by *Nicklow et al.* [2010]). In multiobjective hydrologic model calibration, nondomination sorting partitions candidates model parameterizations into groups that are worse in all objectives ("dominated solutions") and those where their performance is superior to all others in at least one objective ("nondominated solutions"). This partitioning must be done carefully and at a numerical precision that is meaningful. For example, *van Werkhoven et al.* [2009] showed that although trade-off solution sets can be very large (more than 100,000 parameter sets), they may, in fact, represent a very small ranges of performance for component calibration objectives. This implies that the trade-offs themselves are not meaningful.

[5] Building on these observations, this paper poses a basic question: When are multiobjective calibration trade-offs in hydrologic models meaningful? To comprehensively explore this question, we have expanded the four-objective calibration strategy of *van Werkhoven et al.* [2009] focusing on peak flows, low flows, water balance, and flashiness to 392 model parameter estimation experiment (MOPEX) watersheds [*Duan et al.*, 2006] across the United States. Our analysis explores the influence of model structure by analyzing how the multiobjective calibration trade-offs for the lower complexity HYMOD [*Boyle et al.*, 2003; *Moore*, 1985; *Wagener et al.*, 2001] and the moderate complexity HBV [*Bergström*, 1975, 1992, 1995] compare for each of the 392 watersheds. Our results demonstrate that for modern multiobjective calibration frameworks to yield any meaningful measure of model structural failure, users must be able to carefully control the precision by which they evaluate their trade-offs to a level that is reasonably consistent with the variety of uncertainties that exist (e.g., input uncertainty, model structural uncertainty, numerical precision, etc.). Building on *Kollat and Reed* [2007] our study demonstrates that the concept of epsilon-dominance (ε-dominance) [*Laumanns et al.*, 2002] provides an effective means of attaining bounded and meaningful hydrologic model calibration trade-offs. When analyzed at an appropriate precision, our study demonstrates that multiobjective trade-offs are far less frequent than the prior literature has suggested [*Efstratiadis and Koutsoyiannis*, 2010]. However, when trade-offs do exist at a meaningful precision, they do capture structural deficiencies in hydrologic models. This study demonstrates how to use the presence or absence of calibration trade-offs to support hydrologic model selection.

2. Methodology

2.1. Case Study

[6] This study utilizes data from the model parameter estimation experiment (MOPEX) data set [*Duan et al.*, 2006] available from the *National Weather Service* [2011] (available at http://www.nws.noaa.gov/oh/mopex/mo_datasets.htm). The MOPEX data set includes hydrometeorological data for 438 U.S. catchments ranging in size from 67 to 10,329 km^2 located across the conterminous United States. Data available at each catchment includes daily mean areal precipitation (mm), minimum and maximum daily air temperature (°C), daily streamflow (mm), and climatic potential evapotranspiration (mm) based on the NOAA Pan Evaporation Atlas [*Farnsworth and Thompson*, 1982]. Data for many of the MOPEX catchments starts on 1 January 1948 and is available through 31 December 2003. Additional data associated with the 438 basins is available (e.g., soil characteristics, vegetation, etc.), but was not utilized as part of this study. Three hundred ninety-two MOPEX catchments were selected for use in this study based on data availability for a 10-yr calibration period (described in section 3.1). The climatic potential evapotranspiration data available with the MOPEX catchments was replaced with Hamon potential evaporation (PE) [*Hamon*, 1961; *Vorosmarty et al.*, 1998], as this was deemed a more applicable approach. This technique utilized the minimum and maximum daily temperature available with the MOPEX data set, as well as the number of daylight hours per day (which is dependent on the day of the year and the latitude of the catchment), and the saturated vapor pressure (which is estimated based on the daily temperature data available at each catchment). Estimating PE using the Hamon approach resulted in PE estimates that were more consistent with the actual daily temperature at a catchment and did not require additional data, but rather additional calculations based on the available data.

2.2. Rainfall Runoff Models Tested

2.2.1. HYMOD

[7] The lumped conceptual model HYMOD [*Boyle et al.*, 2003; *Wagener et al.*, 2001] (an iteration of the probability distributed model or PDM [*Moore*, 1985]) represents a simple conceptual hydrologic model, which in our study is composed of a snow module, soil-moisture accounting module, and a routing module (see Figure 1). Our HYMOD snow module uses a simple degree-day method [*Bergström*, 1975] for calculating snowmelt. When the average air temperature for a day falls below the temperature threshold for snow (T_t), snow storage occurs. When the average daily air temperature is above the temperature threshold for snowmelt (T_b), snowmelt occurs at the rate defined by the degree-day factor (DDF). The soil-moisture accounting module of HYMOD utilizes a storage capacity distribution function for the storage elements of the catchment. In this module, the storage elements of the catchment are distributed according to a probability density function defined by the maximum soil moisture storage, and the distribution of

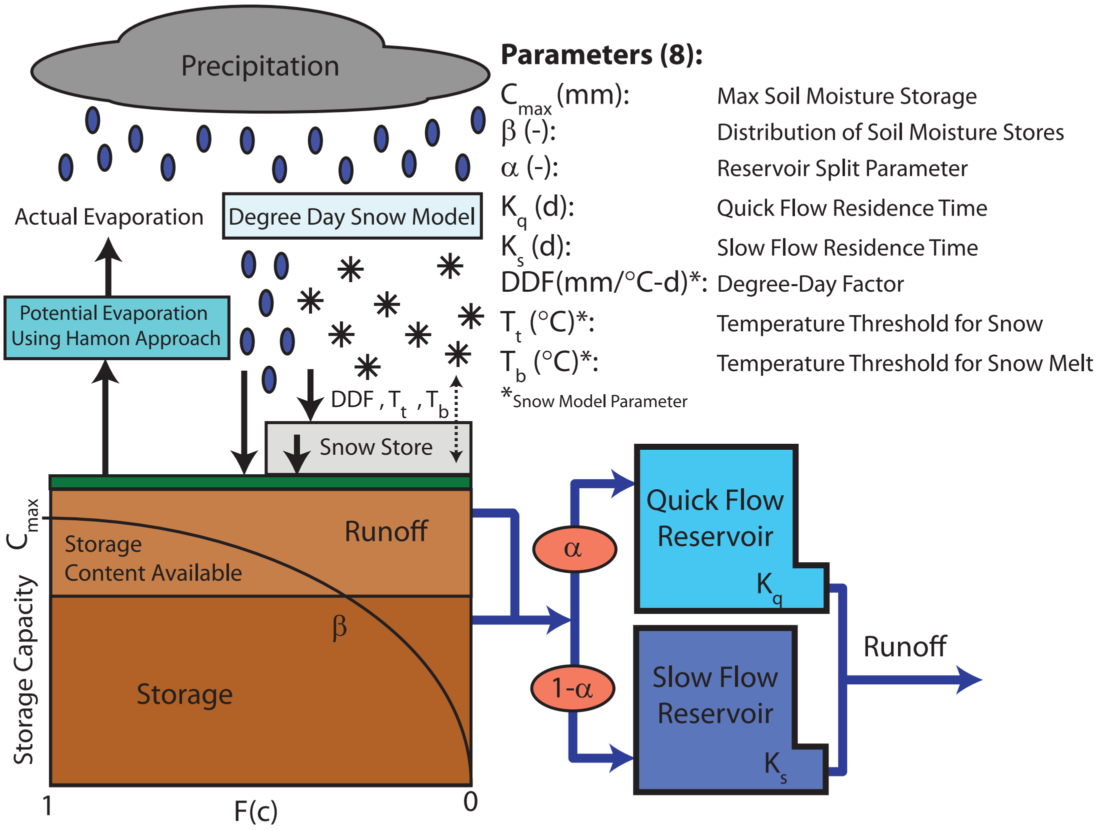
Precipitation
Actual Evaporation
Degree Day Snow Model
Potential Evaporation Using Hamon Approach
DDF , T_t , T_b
Snow Store
C_{max}
Storage Capacity
Storage Content Available
Runoff
β
Storage
1
F(c)
0
α
1-α
Quick Flow Reservoir
K_q
Slow Flow Reservoir
K_s
Runoff
Parameters (8):
C_{max} (mm): Max Soil Moisture Storage
β (-): Distribution of Soil Moisture Stores
α (-): Reservoir Split Parameter
K_q (d): Quick Flow Residence Time
K_s (d): Slow Flow Residence Time
DDF(mm/°C-d)*: Degree-Day Factor
T_t (°C)*: Temperature Threshold for Snow
T_b (°C)*: Temperature Threshold for Snow Melt
*Snow Model Parameter

Figure 1. Diagram of the HYMOD conceptual hydrologic model including the definitions of its eight calibration parameters.

soil moisture stores. The maximum soil moisture storage ($C_{\max}$) represents the capacity of the largest soil moisture store, while the shape parameter (β) describes the degree of spatial variability of the stores [*Wagener et al.*, 2004]. Evaporation from the soil moisture store occurs at the rate of the potential evaporation estimates using the Hamon approach. Following evaporation, the remaining rainfall and snowmelt are used to fill the soil moisture stores. Excess rainfall is sent to the routing module. The routing module divides the excess rainfall using split parameter (α) and routes these through parallel conceptual linear reservoirs meant to simulate the quick and slow flow response of the system. The flow from each reservoir is controlled by the quick flow residence time (K_q) and the slow flow residence time (K_s). The simulated streamflow is therefore the addition of the outputs from each of these reservoirs.

[8] There are a total of eight parameters that must be calibrated for HYMOD as shown in Figure 1, three of which are related to the snow module.

2.2.2. HBV

[9] Our study is focused on a lumped form of the conceptual HBV model [*Bergström*, 1975, 1992, 1995; *Lindström et al.*, 2005; *Seibert*, 2000] that consists of a degree-day snow module, a soil-moisture accounting module, and a runoff response module (see Figure 2). The HBV model represents an increase in complexity relative to the HYMOD model. In fact, HYMOD can be viewed approximately as a simpler form of the more complex HBV model as the various modules are very similar in function, only differing in their parametric complexity. The snow module utilized in HBV operates similarly to the degree-day snow module used in HYMOD, but with an added level of complexity that includes meltwater storage, refreeze, and rain/snow mixing. In order to account for the mix of rain and snow that might occur at air temperatures close to the temperature threshold for snow (T_t), a temperature interval parameter (TTI) is added. The temperature interval parameter specifies temperature bounds where precipitation falls either completely as rain or completely as snow, or a linear mixture of rain and snow for temperatures in between the bounds. The snow store is also assumed to be capable of retaining melt water, expressed as a fraction of its total storage by the water holding capacity of the snow parameter (CWH). The meltwater within the snow can also refreeze according to the refreeze parameter (CFR), which is expressed as a fraction of the degree-day factor. For additional details on the more complex degree-day snow module formulation, refer to *Hamilton et al.* [2000].

[10] The soil-moisture accounting module utilized by HBV is functionally similar to that used by HYMOD. However, instead of assuming that evaporation would occur at the rate of the potential evaporation found using the Hamon approach, the HBV model defines the limiting soil moisture storage at which potential evaporation occurs (LP). For soil moisture storage between 0 and LP, the actual evaporation varies linearly as a fraction of the potential evaporation, and equals the potential evaporation for soil moisture storage at or above LP.

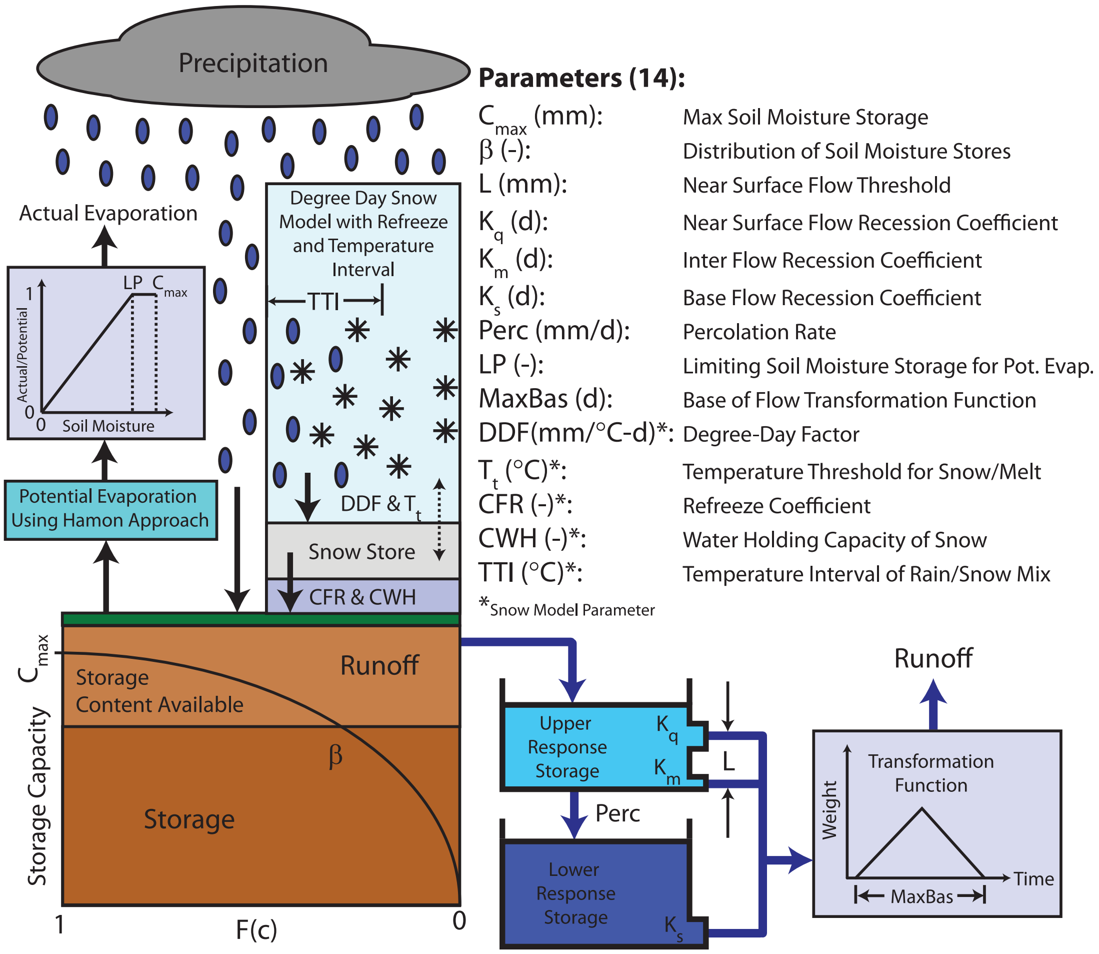
Precipitation
Actual Evaporation
LP C_max
1
0
0
Actual/Potential
Soil Moisture
Potential Evaporation Using Hamon Approach
Degree Day Snow Model with Refreeze and Temperature Interval
TTI
DDF & T_t
Snow Store
CFR & CWH
C_max
Storage Capacity
Storage Content Available
Runoff
β
Storage
1
F(c)
0
Parameters (14):
C_max (mm): Max Soil Moisture Storage
β (-): Distribution of Soil Moisture Stores
L (mm): Near Surface Flow Threshold
K_q (d): Near Surface Flow Recession Coefficient
K_m (d): Inter Flow Recession Coefficient
K_s (d): Base Flow Recession Coefficient
Perc (mm/d): Percolation Rate
LP (-): Limiting Soil Moisture Storage for Pot. Evap.
MaxBas (d): Base of Flow Transformation Function
DDF(mm/°C-d)*: Degree-Day Factor
T_t (°C)*: Temperature Threshold for Snow/Melt
CFR (-)*: Refreeze Coefficient
CWH (-)*: Water Holding Capacity of Snow
TTI (°C)*: Temperature Interval of Rain/Snow Mix
*Snow Model Parameter
Upper Response Storage
K_q
K_m
L
Perc
Lower Response Storage
K_s
Runoff
Transformation Function
Weight
Time
MaxBas

Figure 2. Diagram of the moderately more complex conceptual HBV hydrologic model including definitions of its 14 calibration parameters.

[11] The routing module of the HBV model, similarly to HYMOD, transforms excess rainfall from the soil-moisture storage module to streamflow. The excess rainfall and snowmelt that remains following evaporation, and filling of the soil moisture stores, is routed into an upper response reservoir. Three outlets in the upper response reservoir divide the runoff into near-surface flow, interflow, and percolation to base flow. Flow from the three outlets is defined by the near-surface flow recession coefficient (K_q) the interflow recession coefficient (K_m), and the percolation rate (PERC). A threshold parameter (L) defines the height of runoff in the upper response reservoir at which near-surface flow occurs. Runoff percolating into the lower response reservoir is released according to the base flow recession coefficient (K_s). Runoff released from the upper and lower response reservoirs is then transformed using a triangular distribution with a defined base length (MaxBas).

[12] There are a total of 14 parameters that must be calibrated for the HBV model as shown in Figure 2, five of which are related to the snow module. Again, it is useful to note that HYMOD can actually be viewed as a simpler version of the HBV model as most of its modules are conceptually similar to those in the HBV model. HYMOD and HBV were chosen and formulated in this way to demonstrate how the existence of trade-offs in one or both models is indicative of their structural differences (and/or deficiencies).

2.3. Multiobjective Calibration Objectives

[13] Our multiobjective formulation builds on *van Werkhoven et al.* [2009] to focus on peak flows, low flows, water balance, and flashiness. Adding water balance and flashiness-related signatures to standard error measures provides additional hydrologically relevant information about how watersheds behave and how closely the model matches this behavior [*Sawicz et al.*, 2011; *Yilmaz et al.*, 2008].

2.3.1. Nash-Sutcliffe Efficiency (NSE)

[14] Peak flow errors are emphasized in our use of the Nash-Sutcliffe efficiency (NSE) [*Gupta et al.*, 2009; *Nash and Sutcliffe*, 1970], as the first objective as shown in equation (1),

$$\mathrm{NSE} = 1 - \frac{\sum_{t=1}^{N}\left(Q_{s,t} - Q_{o,t}\right)^2}{\sum_{t=1}^{N}\left(Q_{o,t} - \overline{Q}_o\right)^2}, \qquad (1)$$

where $Q_{s,t}$ is the simulated runoff at time t, $Q_{o,t}$ is the observed runoff at time t, and $\overline{Q}_o$ is the mean observed flow over the calibration period. The summation is performed over $t = 1$ through the number of time steps in the calibration period (N). NSE ranges from 1 (optimal) to $-\infty$, and has been used frequently as a hydrologic model calibration objective.

2.3.2. Transformed Root-Mean-Square Error (TRMSE)

[15] Following prior studies [*Misirli et al.*, 2003; *Tang et al.*, 2006], the second objective emphasizes low flow errors using the Box-Cox transformed [*Box and Cox*, 1964] root-mean-square error (TRMSE) as shown in equation (2),

$$\mathrm{TRMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(\hat{Q}_{s,t} - \hat{Q}_{o,t})^2}, \text{where}\,\hat{Q} = \frac{(1+Q)^{\lambda} - 1}{\lambda}, \quad (2)$$

where $\hat{Q}_{s,t}$ is the Box-Cox transformed simulated runoff at time t and $\hat{Q}_{o,t}$ is the Box-Cox transformed observed runoff at time step t. The summation is performed over time steps 1 through the number of time steps in the calibration period (N). $\hat{Q}$ represents the Box-Cox transformation of the runoff Q, where $\lambda = 0.3$. The Box-Cox transformation, in addition to emphasizing low flow periods, also serves to reduce the impacts of heteroscedasticity in the RMSE calculation.

2.3.3. Runnoff Coefficient Percent Error (ROCE)

[16] The third objective considers water balance by seeking to minimize the average annual runoff coefficient percent error (ROCE) as shown in equation (3),

$$\mathrm{ROCE} = \frac{1}{Y}\sum_{y=1}^{Y}\left|\frac{\overline{Q}_s}{\overline{Q}_o} - 1\right| \times 100\%, \quad (3)$$

where $\overline{Q}_s$ is the mean annual simulated runoff and $\overline{Q}_o$ is the mean annual observed runoff. The summation occurs over years 1 through Y of the calibration period, for which an average annual value is then calculated.

2.3.4. Slope of the Flow Duration Curve (SFDCE)

[17] The fourth objective addresses the flashiness of a watershed's response by minimizing the error in simulating the slope of the flow duration curve (SFDCE) as shown in equation (4),

$$\mathrm{SFDCE} = \left|\frac{Q_{s,67\%} - Q_{s,33\%}}{Q_{o,67\%} - Q_{o,33\%}} - 1\right| \times 100\%, \quad (4)$$

where $Q_{s,67\%}$ is the 67th percentile of the simulated flows, and $Q_{s,33\%}$ is the 33rd percentile of the simulated flows. Likewise, $Q_{o,67\%}$ and $Q_{o,33\%}$ are the 67th and 33rd percentiles of the observed flows. The flow duration curve is the cumulative distribution function of the flows with the flow values plotted on the Y-axis and the probability of exceedance plotted on the X-axis.

2.4. Multiobjective Optimization Algorithm

[18] In this study, the ε-nondominated sorted genetic algorithm II (ε-NSGAII) [*Kollat and Reed*, 2005, 2006; *Reed et al.*, 2007] is used to optimize the model parameters for each of the MOPEX catchments based on the four calibration objectives identified in section 2.3. The ε-NSGAII is a multiobjective evolutionary algorithm (MOEA) [*Coello Coello et al.*, 2007] that uses a population-based search to evolve the Pareto approximate set of hydrologic model calibration parameters (i.e., the parameter sets whose performance in an objective cannot be improved without degrading their performance in one or more other objectives). The ε-NSGAII has been shown to be highly effective at multiobjective calibration of hydrologic models [*Tang et al.*, 2006, 2007; *van Werkhoven et al.*, 2009; P. Reed, D. Hadka, J. Herman, J. R. Kasprzyk, and J. B. Kollat, Evolutionary multiobjective optimization in water resources: The past, present, and future, submitted to *Advances in Water Resources*], and was chosen in this work due to its effectiveness and computational efficiency for the large number of optimization runs required to calibrate all 392 MOPEX catchments. Two unique characteristics of the ε-NSGAII are that it adaptively sizes its evolving population commensurate with search progress, and that it stores the solutions found throughout the run in an ε-dominance [*Laumanns et al.*, 2002] archive. Initially, a small population is used to precondition the search at a low computational cost. As Pareto solutions are found, they are stored in an ε-dominance archive. During later stages of the search, a concept termed "time continuation" [*Goldberg*, 2002] is used to reinvigorate the search by injecting top performing solutions that have been found throughout the run from the ε-dominance archive. The ε-dominance archive stores Pareto solutions found throughout the search according to ε-dominance precision settings specified for each calibration objective. These epsilon settings are critical for performing the nondomination sorting of candidate parameterizations using a meaningful precision as was discussed in section 1.

[19] The ε-NSGAII utilizes ε-dominance archiving in order to prevent the deterioration of the search [*Hanne*, 1999], a phenomenon where MOEAs that utilize fixed populations can lose nondominated solutions from early generations. A key benefit of- ε-dominance archiving is that it allows the user to specify a precision at which to evaluate each of their objectives, which can have dramatic benefits for reducing the computational demands and ensuring numerically meaningful results [*Kollat and Reed*, 2007]. In Figure 3A, a hypothetical example trade-off between NSE and TRSME is shown. This trade-off is also referred to as the Pareto front [*Pareto*, 1896a, 1896b]. However, in most multiobjective model calibrations, it is computationally intractable to find the true Pareto optimal set, so the set found by the MOEA is typically referred to as the Pareto approximate set. In Figure 3A, the trade-off represents the full precision case, meaning that it computes NSE and TRMSE at high levels of precision. Although numerical implementation of the hydrologic model and the error calculations are highly precise (i.e., double precision), users must contemplate at what "significant" precision these calculations are meaningful. We propose that the- ε-dominance concept (beyond its advantages for searches) provides an effective means to understanding and utilizing meaningful precisions in multiobjective calibration problems.

[20] When a meaningful precision is specified according to the ε-precision values for each objective shown in Figure 3A, precision "blocks" are defined where only a single solution is allowed. In Figure 3A, those blocks containing multiple solutions are immediately "thinned" (i.e., the red x's designating eliminated solutions). Only the single solution

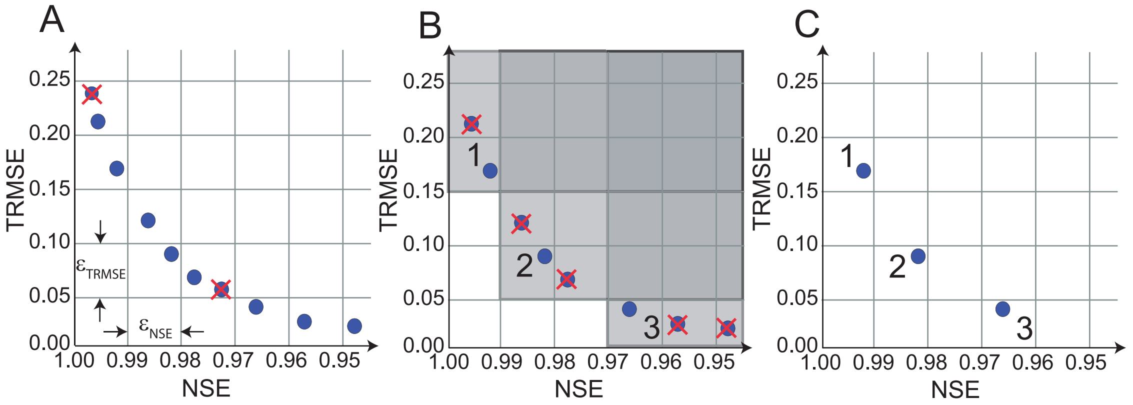

Figure 3. Demonstration of ε-dominance applied to a two-objective calibration problem. A shows the full precision Pareto approximate set and meaningful ε-precision blocks. B shows "block-based" ε-nondomination sorting. C shows the user defined "meaningful" Pareto front based on the specified ε-precision.

closest to the lower-left corner of each epsilon block is allowed to remain. Figure 3B illustrates that the next step is "block-based" ε-nondomination sorting. In Figure 3B, gray shading is used to show the regions that are dominated by solutions 1, 2, and 3. Note that when these solutions-dominated regions intersect, the shading is darker. Solution 1 dominates all of the solutions with the NSE less than 1 and the TRMSE greater than 0.15. Likewise, solution 2 dominates all of the solutions with the NSE less than 0.99 and the TRMSE greater than 0.05. Finally, solution 3 dominates the region where the NSE is less than 0.97 and the TRMSE is greater than 0. The red x's in Figure 3B show which solutions from 3A are eliminated in the block nondomination sorting. Finally, panel C shows the user defined "meaningful" Pareto front based on ε-precision values of NSE = 0.01 and TRMSE = 0.05. These settings mean we are only interested in differentiating NSE and TRMSE at the levels of 0.01 and 0.05, respectively. A key contribution of this is study will be the demonstration of the benefits of using ε-dominance to attain bounded and meaningful hydrologic model calibration trade-offs.

3. Experiment

3.1. MOPEX Data

[21] For this study, 10 yrs of precipitation, temperature, and streamflow data were used to calibrate each of the models. A 1-yr warm-up period was specified, in which the four calibration objectives were not calculated, requiring a total of 11 yr of data for each catchment. Analysis of the 438 available MOPEX catchments [*Duan et al.*, 2006] revealed that 392 of the watersheds had 11 complete years of data from the period 1 October 1961 to 30 September 1972. Consequently, our study considers only the 392 MOPEX catchments that contained complete data for this 11-yr period. The map in Figure 4 shows the 392 MOPEX catchments

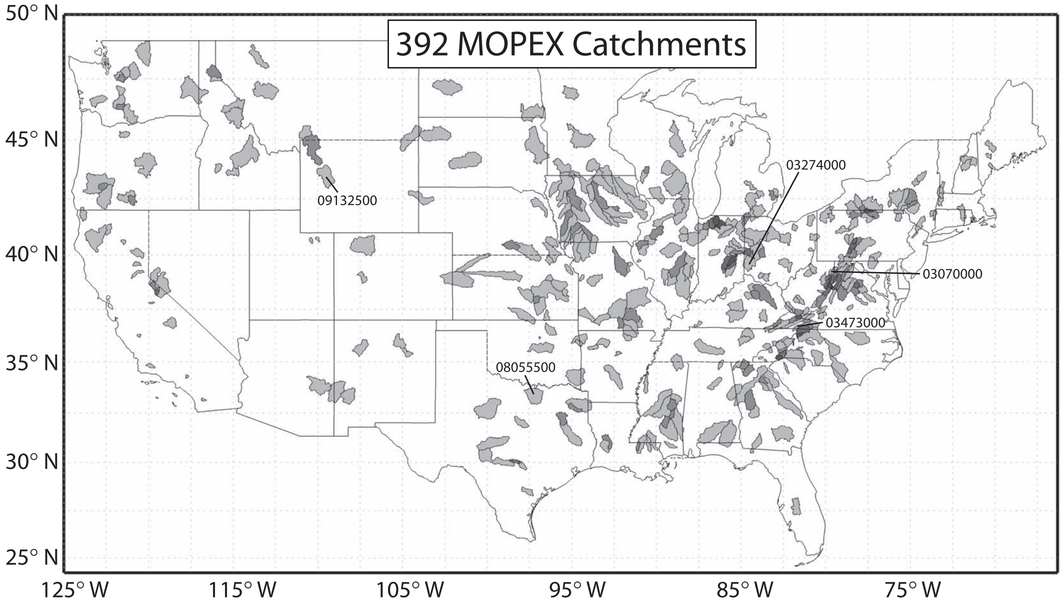

Figure 4. Map of the 392 MOPEX catchments used in this study [*Duan et al.*, 2006]. Labels are provided for five catchments that are explored in detail later in Figure 10 and its associated discussion.

calibrated in this study. Note that many catchments are actually sub-basins of larger catchments (occurring particularly frequently for catchments located in the Appalachian range).

3.2. HYMOD and HBV Calibration Parameters

[22] There are a total of eight calibrated parameters in the HYMOD model (three of which are related to the degree-day snow module). The calibration ranges of each of the HYMOD parameters are shown in Table 1 and are based largely on the maximum range sampled from several recent studies [*de Vos et al.*, 2010; *Kokkonen et al.*, 2006; *McIntyre et al.*, 2005; *Moore*, 2007; *Wagener et al.*, 2004]. In HBV, a total of 14 parameters were calibrated (five of which are related to the degree-day snow module). The calibration ranges of each of the HBV parameters are also shown in Table 1 and are again based on the maximum sampled range from prior studies [*Harlin and Kung*, 1992; *Kokkonen et al.*, 2006; *Lawrence et al.*, 2009; *Liden and Harlin*, 2000; *Ogden et al.*, 2010; *Seibert*, 1997; *Singh*, 2010; *Zhang and Lindstrom*, 1996]. The reader may note that HBV's K_s, K_m, and K_q parameters were allowed to overlap in their sampled ranges, introducing the possibility that the flow components of the HBV model may not be clearly distinguished once calibrated for a specific catchment. However, since we are examining such a large number of catchments (392), we felt it necessary to allow these parameters to vary over an adequate range in order to capture the large range of geographic and climatic conditions present across the conterminous U.S. Additionally, this study does not focus on the resulting parameterizations of the models for each catchment, but rather on the performance of the model simulated streamflow and associated objective values.

3.3. ε-Precision Specification

[23] The ε-precision specified for the NSE objective was specified as $\varepsilon_{\mathrm{NSE}} = 0.01$, representing our interest in performance differences that were at least 0.01 or greater (i.e., 0.99 versus 0.98). This selection was based on numerical analysis that revealed that for high-performing catchments (in terms of NSE), this would capture an error level of at least 1 mm d^{-1} for the streamflow time series. The ε-precision specified for the TRMSE objective was based on analysis of the Box-Cox transformation used in this objective calculation. This analysis revealed that a value of $\varepsilon_{\mathrm{TRMSE}} = 0.0025$ achieved an error level of $\sim$1 mm d^{-1} for the streamflow time series. The ROCE objective was formulated specifically as a percent error (see equation (3)). One percent differences (or $\varepsilon_{\mathrm{ROCE}} = 1.0$) were specified for the ROCE objective. The SFDCE objective was also formulated as a percent error (see equation (4)) with ε-precision specified for 1% differences ($\varepsilon_{\mathrm{SFDCE}} = 1.0$). In general, the ε-precision values of the performance objectives that we have selected for this work are conservative in the sense that they quantify relatively small changes in the objectives relative to the uncertainties present in the system.

Table 1. Summary of Calibration Parameter Ranges Sampled for Both HYMOD and HBV

Parameter (Units)	HYMOD Range	HBV Range
$C_{\max}$ (mm)	0.0–2000.0	0.0–2000.0
LP (-)	–	0.3–1.0
β(-)	0.0–7.0	0.0–7.0
α(-)	0.0–1.0	–
K_q (d)	1.0–7.0	0.5–20.0
K_m (d)	–	1.0–100.0
K_s (d)	7.0–20,000	10.0–20,000
L (mm)	–	0.0–100.0
PERC (mm d^{-1})	–	0.0–100.0
MaxBas (d)	–	24.0–120.0
DDF (mm °C^{-1} d^{-1})	0.0–20.0	0.0–20.0
T_t (°C)	−3.0–3.0	−3.0–3.0
T_b (°C)	−3.0–3.0	–
CFR (-)	–	0.0–1.0
CWH (-)	–	0.0–0.8
TTI (°C)	–	0.0– 7.0

3.4. Parameterization of the ε-NSGAII

[24] Parameters related to the ε-NSGAII algorithm include an Simulated Binary Crossover (SBX) probability of 100% with an SBX distribution index of 15 [*Deb and Agrawal*, 1995]. The polynomial mutation probability was specified differently for each the HYMOD and HBV calibration runs using the rule-of-thumb $1/X$, where X is the number of real coded variables [*Deb*, 2001]. Hence, the mutation probability for the HYMOD runs was set to 12.5% and for the HBV the runs were set to 7.14%, both with a distribution index of 20.

[25] The population size of the ε-NSGAII was initially set at 12 individuals, and was permitted to grow to an upper bound of 10,000 individuals. Run length was set to 50 generations per run, whereby after 50 generations of evolution, 25% of the ε-dominance archive was injected into a subsequent larger population (composed of 75% new random solutions). To prevent preconvergence for catchments that exhibited significantly collapsed trade-offs (i.e., there may exist only one solution in the collapse trade-off), a lower bound of 100 individuals was specified following the first run (or the first 50 generations of evolution), so that a sufficient population size could be used to evolve the calibration trade-off. Each model and catchment was run for a total duration of one million evaluations per run over 50 random seed trials (to account for variability in the initialization of the ε-NSGAII). This represents a total of 50 million evaluations for each of the 392 MOPEX catchments for each HYMOD and HBV.

[26] In total, the runs for this experiment encompassed 39.2 billion model evaluations (i.e., model simulations for the 10 yr calibration period) and were performed on Pennsylvania State University's CyberStar cluster. The CyberStar cluster is composed of 2048 processing cores (1536 quad-core Intel Nehalem 2.66 GHz processors and 512 quad-core AMD Shanghai processors), each with access to at least 3GB of RAM.

4. Results and Discussion

[27] Following the calibration runs for each of the 392 MOPEX catchments, results from all 50 random seed trials were combined and ε-nondomination sorting was performed over the 50 trials to create a "best-known" Pareto approximate set for each model and each catchment. The Pareto approximate sets reported throughout the remainder of

section 4 reflect these best-known trade-offs resulting from a total of 50-million model simulations for each catchment.

4.1. What is a Meaningful Trade-Off?

[28] In order to illustrate how ε-dominance is critical to attaining bounded and meaningful hydrologic model calibration trade-offs, in Figure 5 we demonstrate the application of ε-dominance to a commonly studied catchment, the Leaf River near Collins, Mississippi, which has frequently been shown in the literature to exhibit trade-offs between many of the common calibration objectives [*de Vos and Rientjes*, 2008; *Gupta et al.*, 1998; *Tang et al.*,

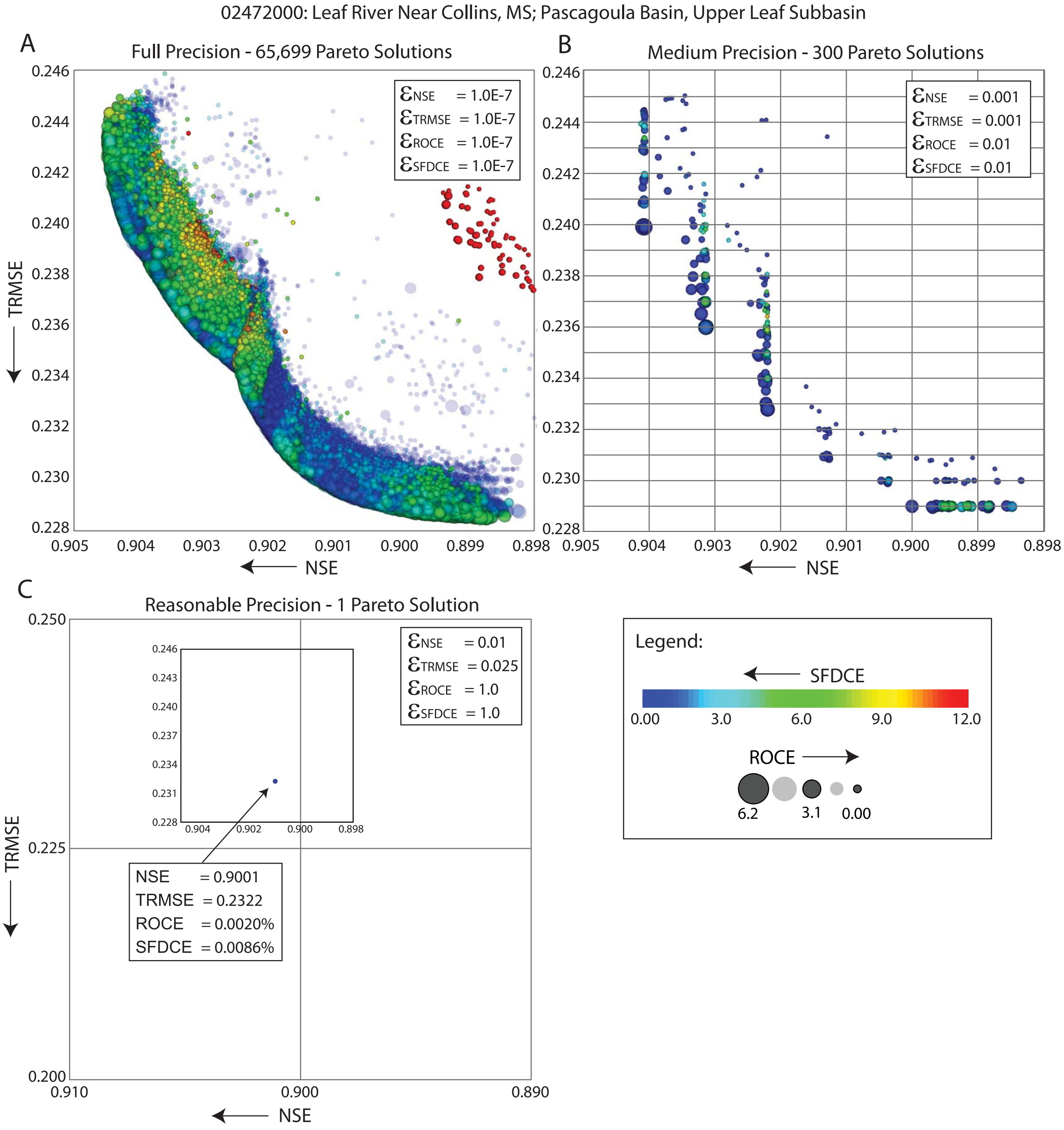

Figure 5. Demonstration of ε-dominance applied to HBV calibration results for the Leaf River catchment near Collins, Mississippi. A shows the "full precision" version of the Pareto approximate parameter set with NSE and TRMSE plotted on the *X* and *Y* axes, SFDCE plotted as color, and ROCE plotted using the size of the markers (arrows point toward the preferred region of the space). B demonstrates how the ε-precision settings shown in the top right corner of the panel applied to each of the performance objectives results in 300 Pareto approximate parameter sets. C demonstrates how these ε-precision settings can be further refined (to the meaningful precision defined in section 3.3) to result in a single, high-performing Pareto approximate solution.

2006; *Vrugt et al.*, 2003; *Yapo et al.*, 1998]. In each of the plots, NSE and TRMSE are plotted on the *X*- and *Y*-axes. SFDCE is plotted using color where blue represents low SFDCE and red high SFDCE. ROCE is plotted using the size of the markers where small markers represent low ROCE and large markers represent large ROCE. In Figure 5A, the multiobjective calibration results attained using HBV and the ε-NSGAII with the parameter settings described in section 3.4, and highly precise ε-precision values of 1.0E-7 for each of the calibration objectives are shown. Note that these precisions represent commonly employed values in the prior literature [*Gupta et al.*, 1998; *Tang et al.*, 2006, 2007; *van Werkhoven et al.*, 2009; *Yapo et al.*, 1998]. These "full-precision" settings result in 65,699 Pareto approximate parameter sets, similar to results shown by *van Werkhoven et al.* [2009]. Moreover, careful examination of the relative ranges of each of the calibration objectives reveals that the trade-off shown may not actually be meaningful (e.g., NSE ranges between 0.898 and 0.905 in Figure 5A).

[29] Figure 5B shows what the full precision calibration problem shown in Figure 5A would look like if precision settings of NSE = 0.001, TRMSE = 0.001, ROCE = 0.01, and SFDCE = 0.01 were applied to each of the calibration objectives. The grid associated with the NSE and TRMSE ε-precision values is shown within the plot in Figure 5B. In this case, the full precision set is reduced to 300 Pareto approximate parameter sets. However, this still represents a level of precision that is more precise than is necessary given the uncertainty present in the observations and model. In other words, we need to ask the question: Is resolving NSE to the 0.001 level in the presence of the other objectives meaningful?

[30] Figure 5C shows what the full precision multiobjective calibration problem shown in Figure 5A would look like if precision settings of $\varepsilon_{\mathrm{NSE}} = 0.01$, $\varepsilon_{\mathrm{TRMSE}} = 0.025$, $\varepsilon_{\mathrm{ROCE}} = 1.0$, and $\varepsilon_{\mathrm{SFDCE}} = 1.0$ were applied to each of the calibration objectives (i.e., the settings established for this study in section 3.3). These settings correspond with resolving the relative percent error in ROCE and SFDCE to the 1% level. In addition, we could think of the epsilon applied to the NSE objective as resolving NSE to levels of 0.90, 0.91, 0.92, etc. In this case, we have to "zoom out" from the original plot in Figure 5C to show what this ε-precision actually represents (see the large grid in Figure 5C). When we do this, we find that given these "meaningful precision" settings, the original set containing 65,699 Pareto approximate solutions collapses to a single solution with NSE = 0.9001%, TRMSE = 0.2322%, ROCE = 0.0020%, and SFDCE = 0.0086%. This represents an extremely high-performing parameter combination, but eliminates the other 65,698 solutions that are representative of an overly precise quantification of the calibration objectives. In terms of model performance, Figure 5C designates a full collapse of the four-objective problem to its minimum geometry (i.e., a zero dimensional single point satisfying all objectives). It could be argued that the ε-precision values used in this study may still be too precise.

[31] Figure 6 shows the actual calibrated simulation results associated with each of the Pareto approximate sets shown in Figures 5A–5C. Recall that each individual marker on the plots in Figure 5 represents a Pareto approximate model parameter set for HBV. Figure 6 shows a portion of the runoff time series for the Leaf River catchment during the calibration year 1967 with the observed runoff shown in blue, and the precipitation shown as black bars at the top of the plot. The range of the simulations produced by the 65,699 Pareto approximate parameter sets associated with the "full precision" multiobjective calibration shown in Figure 5A is shown in pink. The range of the 300 simulations produced using the medium precision settings from Figure 5B is shown in light red. Finally, the single-calibration parameter set associated with the ε-precision settings used in this study (see Figure 5C) is shown using a dark red line. This figure demonstrates that while there are 65,699 Pareto approximate parameter sets in the full precision calibration case, the range of performance that these parameter sets actually produces is very small (illustrated by the narrow pink band in the plot). Additionally, the simulation produced using the single parameter set generated using a reasonable ε-precision setting performs very well, and would likely be of significant interest to the modeler without the confounding information of the remaining 65,698 solutions as well as the severe computational challenge posed by the full precision multiobjective calibration [*Kollat and Reed*, 2007].

4.2. Meaningful ε-Precision and Trade-Off Collapse

[32] Recall that ideally, multiobjective trade-offs should not exist for hydrologic models. Figure 7 demonstrates

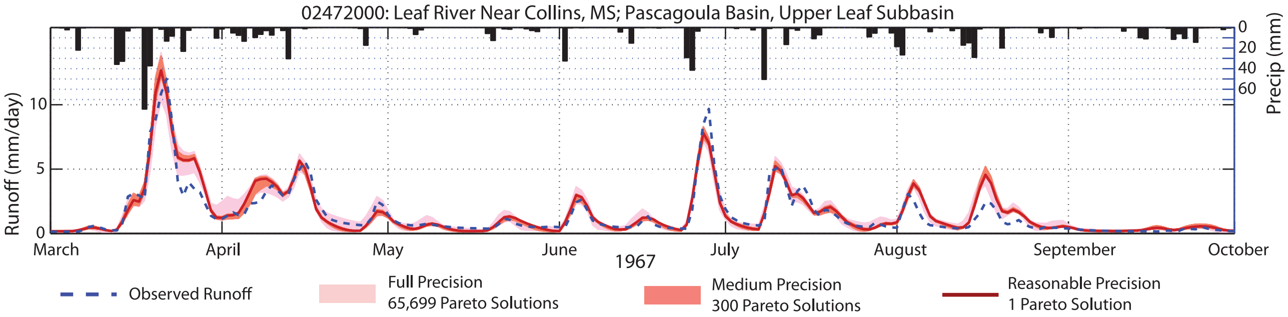

Figure 6. Streamflow plot showing the simulations associated with the full (pink), medium (light red), and reasonable precision (dark red) Pareto approximate sets from Figures 5A, 5B, and 5C along with the observed runoff time series (blue dashed line) and precipitation (black bars) for 8 months during calibration year 1967.

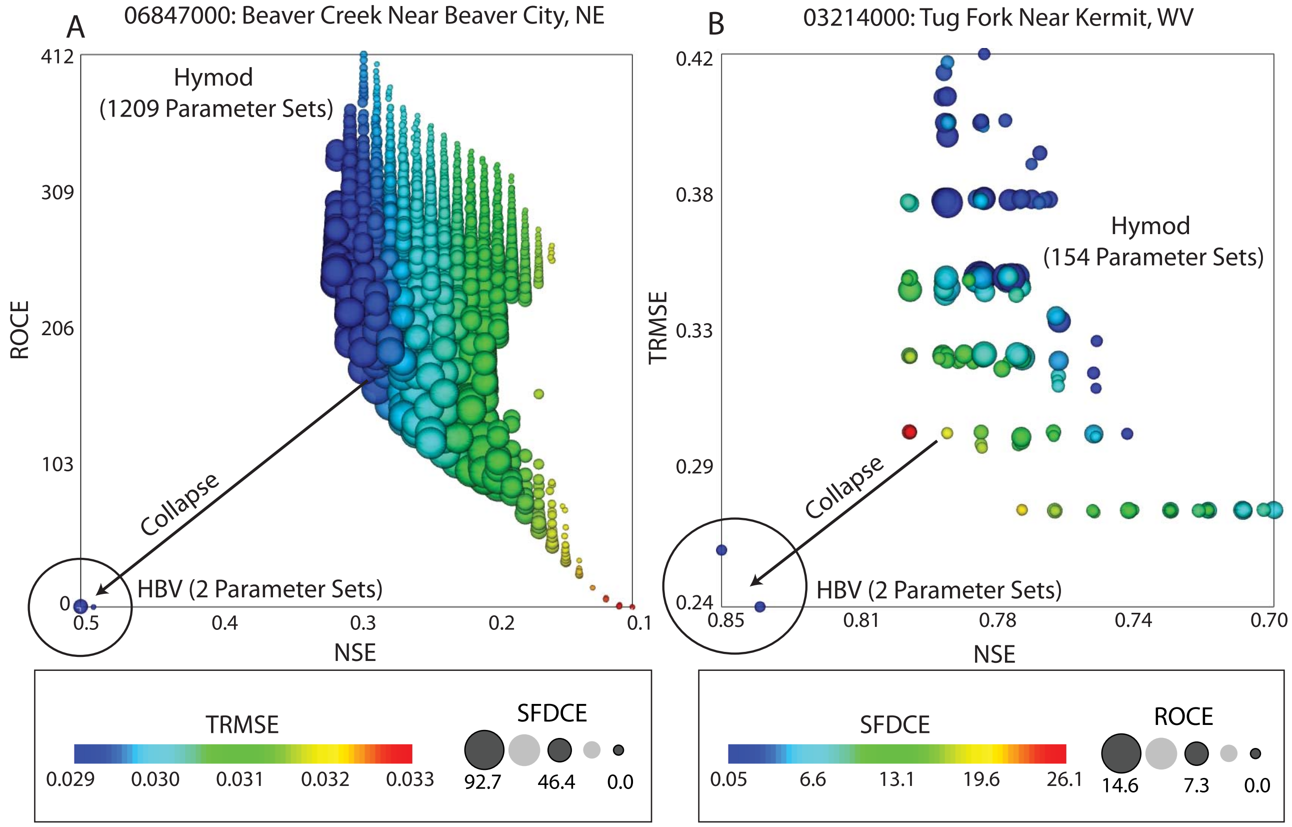

Figure 7. Plots demonstrating the Pareto approximate parameter set collapse resulting from both the models' abilities to simulate the streamflow, and the application of reasonable ε-precision to the calibration objectives. A shows the Pareto approximate parameter sets for HYMOD and HBV (circled) applied to the Beaver Creek, Nebraska catchment with NSE and ROCE plotted on the X and Y axes, TRMSE plotted as color, and SFDCE plotted using the size of the markers. B shows the Pareto approximate sets for HYMOD and HBV (circled) for the Tug Fork, West Virginia catchment where NSE and TRMSE are plotted on the X and Y axes, SFDCE plotted as color, and ROCE plotted using the size of the markers. Arrows in both panels point toward the preferred region of the space.

these points by showing how reasonable ε-precision applied to the multiobjective calibration objectives results in a collapse of the calibration trade-offs with slight increases in model complexity. In many catchments, the additional degrees of freedom (additional model parameters) present in the HBV model allow it to outperform the simpler HYMOD model. Figure 7A shows the HYMOD and HBV Pareto approximate parameter sets for Beaver Creek near Beaver City, Nebraska with NSE and ROCE plotted on the X- and Y-axes, TRMSE plotted using color, and SFDCE plotted using the size of the markers. Here the increased complexity of the HBV model clearly allows it to fit more accurately the runoff time series for this catchment. Additionally, reasonable ε-precision values facilitate the collapse of the HBV trade-off due to its increased ability to satisfy the full suite of high flow, low flow, water balance, and flashiness objectives. In Figure 7A, only two Pareto approximate parameter sets exist at this level precision. The two HBV parameterizations themselves are not strongly distinguishable and actually could be considered analogous to one another. Figure 7B similarly shows the HYMOD and HBV Pareto approximate parameters sets for a different catchment, Tug Fork near Kermit, West Virginia, but with NSE and TRMSE plotted on the X- and Y-axes, SFDCE plotted using color, and ROCE plotted using size. Figure 7B further illustrates the structural collapse afforded by the more complex HBV model. Both Figures 7A and 7B clearly illustrate improved model identification with the slight increase in model complexity when transitioning from HYMOD to HBV. Epsilon-dominance is well-suited for capturing this dimensional collapse of trade-offs, which would be hidden in numerical noise when an unnecessarily high level of precision is utilized. Epsilon-dominance ultimately provides a straightforward conceptual approach for assessing dimensional collapse in the resulting Pareto approximate sets for HBV and HYMOD.

[33] This type of analysis can be extended across the calibration results of all 392 MOPEX catchments as shown in Figure 8. Figures 8A and 8B show maps of the Pareto approximate parameter set sizes generated by HYMOD and HBV for each of the 392 MOPEX catchments. The coloring of the catchments indicates the Pareto approximate parameter set size with red indicating more than 1000 Pareto approximate parameter sets and blue indicating complete

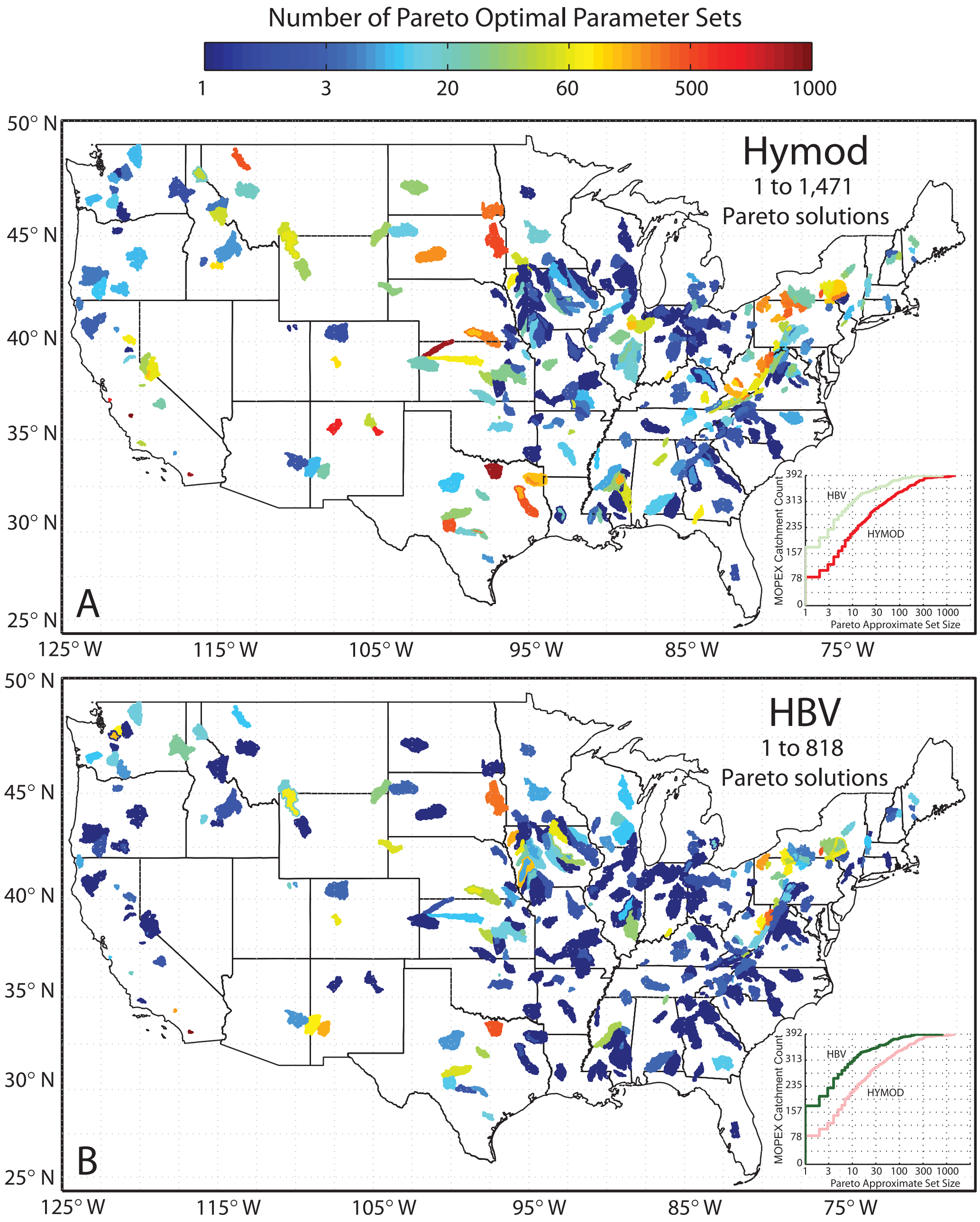

Figure 8. Maps of Pareto approximate parameter set sizes for the 392 MOPEX catchments generated by both HYMOD and HBV. Red represents large Pareto sets (1000+ parameter sets), while blue represents small (or collapsed) Pareto sets (one optimal parameter set). A cumulative distribution function plot of the MOPEX catchment count versus Pareto approximate set size for HYMOD and HBV is shown in the corner of each map.

collapse to a single Pareto approximate parameter set. The calibration results for HYMOD ranged between 1 and 1471, and for HBV between 1 and 818 Pareto approximate parameter combinations in each set. Also shown on each map is a cumulative distribution function (CDF) plot of MOPEX catchment count versus Pareto approximate set size for both HYMOD and HBV. The most notable contrast between Figures 8A and 8B is that the HBV model clearly exhibits more collapse in its Pareto approximate parameter sets (shown both in terms of the maps, and the CDFs), indicating that HBV's additional complexity is typically beneficial in more accurately reproducing the streamflow of each catchment. However, there is clearly a geographical influence in the performance of both models, as it appears to be particularly difficult for both HYMOD and HBV to simulate catchments along parts of the Appalachian Range (i.e., the Pareto approximate set sizes in this region tend to be large).

[34] Additional analysis was performed by combining the Pareto approximate parameter sets of both HYMOD

and HBV to examine the relative contributions of each model to the "global Pareto approximate parameter set" produced by combining the results from both models for each catchment. This analysis concluded that of the global Pareto approximate parameter sets generated for the 392 MOPEX catchments, there are only 24 (6.1%) where HYMOD actually contributes to the best known Pareto approximate sets and of these, there are only 17 (4.3%) where HYMOD contributes more Pareto approximate parameter sets than HBV. Additionally, for the 73 catchments (or 18.6% of the MOPEX catchments), where HYMOD's Pareto approximate parameter set size is smaller than HBV's, HBV results actually dominate the collapsed HYMOD sets in all but nine catchments (2.3%). In these nine catchments, there tends to be a mixed contribution from both HYMOD and HBV, indicating that both models are attaining highly similar levels of performance. In cases where both HBV and HYMOD made significant contributions to their joint Pareto approximate set, the models are largely equivalent in their performance.

[35] Figure 9 provides cumulative distributions functions (CDFs) for both the HYMOD and HBV models across their calibration objectives (Figure 9A) and their parameters (Figure 9B). The CDF for HYMOD is shown as a dashed red line and for HBV as a solid green line. These CDFs were produced by aggregating the Pareto approximate parameters sets generated by each model across all 392 MOPEX catchments. This represents a sample size of 20,640 Pareto approximate parameter sets for HYMOD and 5639 Pareto approximate parameter sets for HBV (the sample size is obtained by summing the Pareto set sizes on the maps shown in Figure 8). For the calibration objectives, the preferred or optimal side of the plot is always shown to the left (note that although NSE is maximized, the *X*-axis for this plot increases from right to left).

[36] In terms of performance across the calibration objectives, HBV generally outperforms HYMOD most in terms of NSE (see Figure 9A), although it tends to slightly outperform HYMOD in ROCE and SFDCE as well. The distributions for TRMSE are similar across both models. The distributions of the model parameters shown in Figure 9B for both HYMOD and HBV are indicative of the applicability of the sampled range of each parameter. For many parameters, we see that they occur throughout their sampled range across the 392 catchments. Notable deviations to this include the percolation parameter (PERC) for the HBV model, which tends toward lower values (<25%) for most catchments. Interestingly, for most parameters that have been identified as functionally similar across HYMOD and HBV, their CDFs appear similar, confirming the stance that HYMOD represents essentially a subset of the more complex HBV structure. The only deviation to this is for the parameter β, which suggests that the way in which model components are connected to other parts of a model can have an impact on their functional role in the model (see discussion by *Clark et al.* [2008]). Since some of the modules (e.g., evaporation, runoff, etc.) that make up the overall structures of HYMOD and HBV differ from one another, the functional role of the soil moisture module (which utilizes the β parameter) may differ between the two even though it is structurally similar, especially when compared over a large number of catchments.

4.3. Precision and Trade-Off Collapse: What Does it All Mean?

[37] Figure 10 provides detailed streamflow plots for five examples of MOPEX catchments for the period October 1969 through September 1971 of the 10-yr calibration. The observed runoff is shown as a dashed blue curve, the precipitation as black bars, and the Pareto approximate parameter set simulations associated with HYMOD and HBV are shown in red and green, respectively. Figures 10A through 10C show catchments where the trade-offs do not collapse for either model (i.e., both models exhibit difficulty simulating the streamflow). Figures 10D and 10E show catchments where the trade-offs collapse for both models due to generally good performance.

[38] In Figure 10A, both models exhibit difficulty modeling the Cheat River located at Rowlesburg, West Virginia (see gage 03070000 in Figure 4). HBV's trade-off is composed of 259 parameter sets and HYMOD's is composed of 174. Careful examination of the plot reveals "peaking" behavior in both models that is simply not present in the observed streamflow. The Cheat River Basin is characterized by the existence of maze caves and single conduit caves [*Springer et al.*, 1997], indicating that complex subsurface flow conditions likely exist that neither HYMOD nor HBV are equipped to model. In Figure 10B, we show a snowmelt dominated catchment located on the North Fork Gunnison River near Somerset, Colorado (see gage 09132500 in Figure 4). This catchment is entirely dominated by spring snowmelt that neither HYMOD's nor HBV's simple degree-day snow model is sufficient to capture. These simple snow modules fail to capture the complex elevation-dependent snowmelt scenarios that occur in this catchment. As a result, both models exhibit significant trade-offs due to their inability to accurately represent the system. In fact, the performance of both HYMOD and HBV at many snowmelt-dominated catchments in mountainous regions of the U.S. are similarly characterized by large trade-offs, while their performance in lower elevation regions, where snow occurs, is generally much improved (i.e., collapsed). This is likely reflective of their inability to effectively model the complexities that exist in elevation-dependent snowmelt release timing and magnitude. Figure 10C shows a catchment located at the Elm Fork Trinity River near Carrollton, Texas that is clearly influenced by reservoir releases (see gage 08055500 in Figure 4). This gage is located downstream from the Lewisville Dam, resulting in a highly regulated flow at the gage due to reservoir operations. This example demonstrates that, while the increased complexity of the HBV model greatly improves its ability to reproduce the streamflow (HBV's Pareto approximate trade-offs collapse to 208 parameter sets while HYMOD's trade-offs remain at 1231 parameter sets), it is still fundamentally lacking model structural components that would allow it to more accurately reproduce the runoff given this anthropogenic disturbance (see Figure 4).

[39] Alternatively, Figure 10D provides an example of a catchment where both HYMOD and HBV achieve an excellent fit, and hence their Pareto approximate calibration trade-offs collapse to a single solution. The Great Miami River at Hamilton, Ohio (see gage 03274000 in Figure 4) is minimally regulated at low flow, and contains five retarding basins upstream of the gage to control flood flow

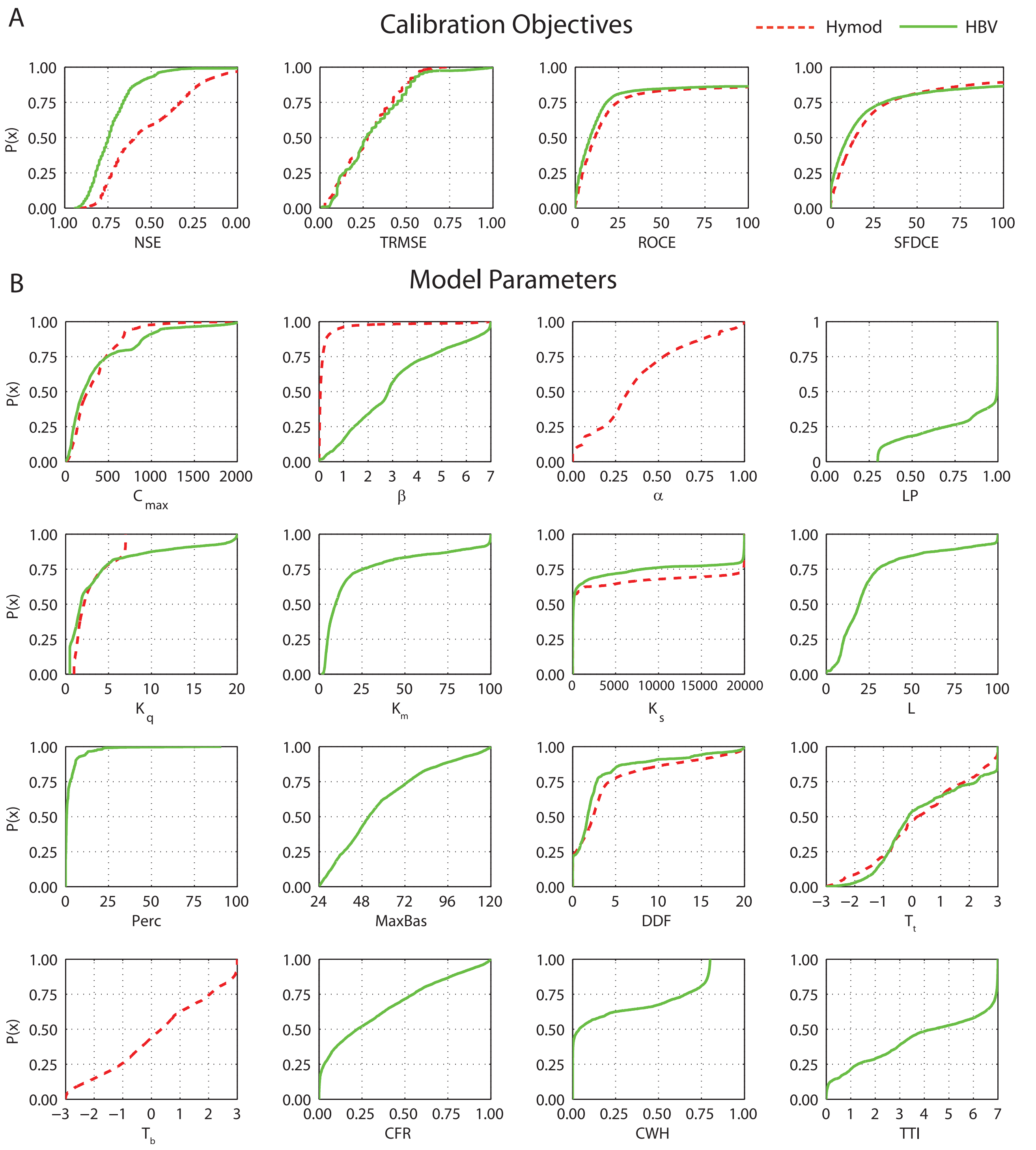

Figure 9. Cumulative distribution functions (CDFs) of the objectives and model parameters across all Pareto approximate parameter sets generated for the 392 MOPEX catchments. CDFs for HYMOD are shown as red dashed lines and for HBV are shown as solid green lines. The full range of the ROCE and SFDCE CDFs were trimmed to 100% error for visualization purposes.

[*U.S. Geological Survey (USGS)*, 2010]. The geology, climate, and nature of the anthropogenic influence on this catchment apparently combine to form ideal modeling conditions for both HYMOD and HBV. Figure 10E is meant to provide a contrast to the Cheat River shown in 10A, as the South Fork Holston River near Damscas, Virginia (see gage 03473000 in Figure 4) is located southwest of the Cheat River, and at the southern end of the Appalachian Range. It is clear from this figure that the subsurface flow issues present at the Cheat River are not shared by this catchment.

[40] Solving the multiobjective calibration problem using four calibration objectives inherently allows for the

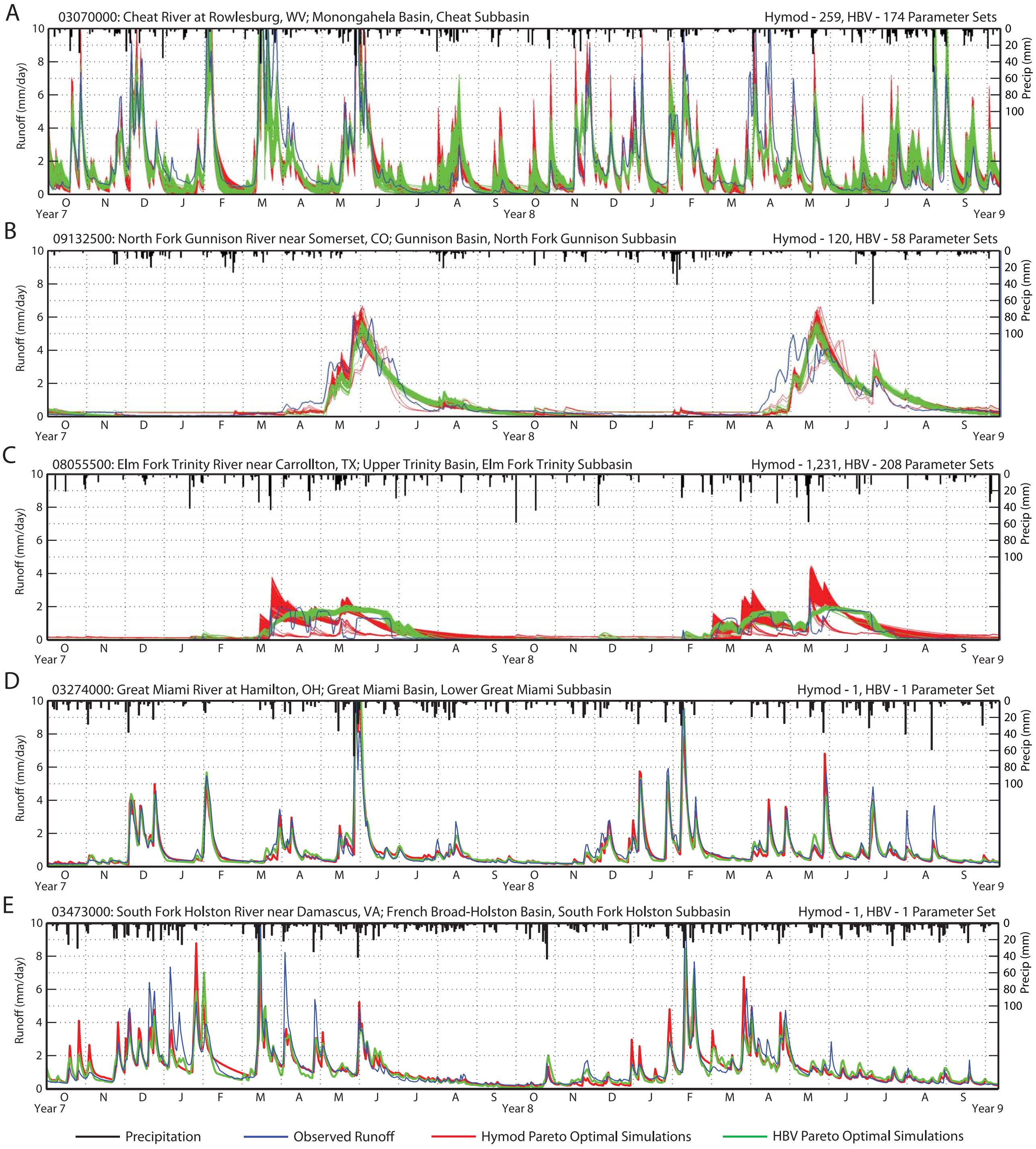

Figure 10. Detailed streamflow plots for five MOPEX catchments. Each plot displays the observed time series in blue, precipitation as black bars, and the simulations associated with HYMOD's and HBV's Pareto approximate parameter sets in red and green, respectively (for the period October 1969 through September 1971). A shows a groundwater dominated catchment (the Cheat River, WV) that is extremely difficult for both models. B shows a snowmelt dominated catchment in Colorado. C shows an anthropogenically impacted catchment in Texas. D and E show catchments in OH and VA, respectively, for which HYMOD and HBV both perform reasonably well resulting in collapse of their Pareto approximate parameter sets.

analysis of any of the subproblems from which it is composed. To illustrate, while we optimized the calibration problem using four objectives, we have at the same time also solved four three-objective calibration problems, six two-objective calibration problems, and four single-objective problems, for a total of 15 problems. Beginning with the Pareto preference ordering work of *Das* [1999], and extending to the work by *Khu and Madsen* [2005] and *di Pierro* [2006], it has been shown that exploring the subproblems within the larger multiobjective problem can be of value in identifying the significance of the conflicts that exist as well as determining which solutions are optimal in the most subproblems. For example, multiobjective calibration subproblems that include the NSE objective may result in larger trade-offs for some catchments, while calibration subproblems that include the SFDCE objective may result in larger trade-offs for others. This results in a changing degree of conflict that exists between pairs of objectives across individual catchments that differ in their controlling processes.

[41] For each of the subproblems contained within the full four-objective calibration problem, we now analyze where the most significant conflicts exist in order to identify meaningful patterns. To accomplish this, ε-nondomination sorting was performed on the moderately large (20 or more) Pareto approximate parameter sets generated by HBV using the ε-precision settings established in section 3.3 to focus on the presence or absence of trade-offs for each of the 11 subproblems within the full four objective calibration problem (i.e., there are no trade-offs associated with the four single-objective problems). Hierarchical clustering using a Euclidean distance metric [*Hastie et al.*, 2009] was then performed according to the Pareto approximate parameter set sizes across the 11 subproblems. For the clustering analysis, the Pareto approximate set sizes of each subproblem were standardized to mean zero and a standard deviation of 1.

[42] Figure 11A shows a "heat map" of the clustering of the 52 catchments where HBV identified four-objective trade-offs containing 20 or more Pareto approximate parameter sets. The rows of the heat map represent each of the 52 catchments and the columns of the heat map represent the 11 subproblems as labeled. The coloring of the heat map represents the standardized Pareto set size, where red indicated large or noncollapsed sets and blue indicates small or relatively collapsed sets. The results of the cluster analysis revealed the existence of two primary clusters as denoted in Figure 11A. We define the "blue cluster" as being the one characterized by large trade-offs in the NSE-TRMSE-ROCE combination of objectives and limited trade-offs in the NSE-ROCE-SFDCE combination (refer to the dashed boxes in the "blue cluster" portion of the heat map). We define the "red cluster" as being characterized by large trade-offs in the NSE-ROCE-SFDCE combination, but small trade-offs in the NSE-TRMSE-ROCE combination (again refer to the dashed boxes). The red and blue clusters represent the first level of clustering into two groups for the hierarchical clustering analysis.

[43] Figure 11B maps the catchments associated with each of the red and blue clusters, where we can see that the red cluster members occur predominantly in the Midwest and the blue cluster members occur predominantly around the Appalachian Range. Previously, in Figure 10A we identified the Cheat River in West Virginia as exhibiting significant surface to subsurface flow interactions where both models failed to reproduce the peaking behavior of the observed streamflow. This mode of failure would result in severe water balance errors and this is consistently what we see in the blue cluster catchments (larger trade-offs in subproblems, including ROCE). However, in the red cluster catchments located throughout the Midwest, we observe more difficulty in fitting the "flashiness" of the time series, as the subproblems that include the SFDCE objective tend to exhibit larger trade-offs. Recent analysis of the MOPEX catchments conducted by *Wang and Hejazi* [2011] suggests that the red cluster catchments predominantly occur in regions heavily impacted by agricultural use. Overall, Figure 11 supports our contention that the presence and absence of trade-offs can be used as a diagnostic for model selection and the identification of structural deficiencies.

5. Conclusions

[44] In this study, we applied a four-objective calibration strategy focusing on peak flows (Nash-Sutcliffe efficiency), low flows (Box-Cox transformed root-mean-square error), water balance (runoff coefficient error), and flashiness (slope of the flow duration curve error) to 392 model parameter estimation experiment (MOPEX) watersheds across the United States using the relatively simple HYMOD, and the slightly more complex HBV hydrologic models, both of which are widely used. The calibration runs were conducted for each model and each catchment by aggregating 50-million total evaluations using the ε-dominance nondominated sorted genetic algorithm II (ε-NSGAII). Our analysis was designed to answer the following key question: When are multiobjective calibration trade-offs in hydrologic models meaningful?

[45] First, in addressing this question we have shown that the block-based nondomination sorting implicit to ε-dominance is critical to attaining bounded and meaningful hydrologic model calibration trade-offs as was demonstrated on the commonly examined Leaf River catchment near Collins, Mississippi. Here we showed that traditional nondomination sorting using highly precise error calculations yields a severe growth of the Pareto approximate parameter sets (65,000 in the Leaf River example), but the trade-offs had extremely small effective ranges for each of the calibration objectives. The Leaf River Pareto approximate set "collapsed" to a single optimal solution using "meaningful" ε-precision for each calibration objective. This example suggests that when calibrating at an appropriate precision, multiobjective trade-offs are far less frequent than prior literature has suggested. In fact, for the MOPEX catchments, 80% of HBV's and 55% of HYMOD's Pareto approximate parameter sets collapse to 10 or fewer parameter sets when using the ε-precision settings of this study. In the majority of cases, the trade-offs probably are not meaningful and reflect a significant dimensional collapse relative to the theoretical potential for a large, four-objective Pareto front geometry.

[46] When trade-offs do exist at meaningful precision, we have demonstrated that they can be used as a diagnostic for model selection and for assessing structural failures in models. Our analysis explored the influence of model structure

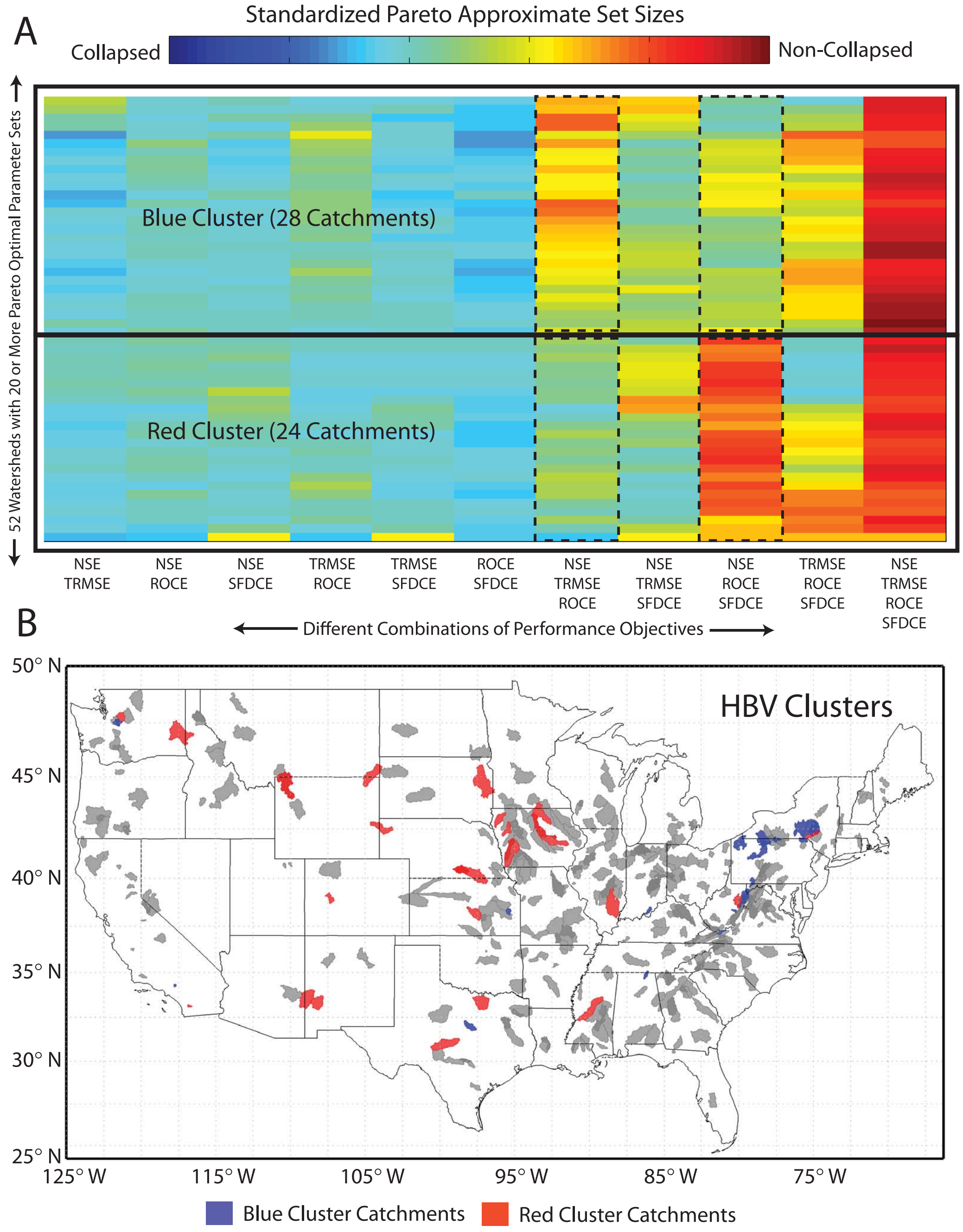

Figure 11. Hierarchical clustering analysis of the 52 catchments for which HBV generated Pareto approximate parameter sets with 20 or more solutions. A shows the first level clustering of these catchments based on the 11 subproblems with the rows being the 52 catchments, the columns being the 11 subproblems, and the coloring is the standardized Pareto set size for each subproblem. B shows a map of the blue and red cluster catchments from Figure 11A.

by analyzing how the multiobjective calibration trade-offs for the lower complexity HYMOD and the moderately higher-complexity HBV hydrologic models compared over the 392 MOPEX catchments. We showed several detailed cases where the additional complexity of HBV was sufficient to produce a trade-off collapse to a well-identified single solution, whereas the simpler HYMOD was unable to achieve similar performance. Additional detailed analysis of both catchments exhibiting collapsed and noncollapsed trade-off surfaces revealed that in many cases, model structural failings could be easily attributable to factors such as groundwater interactions, snowmelt scenarios, and anthropogenic influences. In summary, multiobjective calibration trade-offs should be the exception and not the rule when they are evaluated at an appropriate precision for multiple candidate hydrologic models across diverse hydro-climatic conditions.

[47] At first sight, the methodology proposed in this paper might seem to be opposing the generally accepted problem

of equifinality [*Beven*, 2006; *Beven et al.*, 2011]. Our approach collapses similarly performing calibration parameter sets to a single representative one, while approaches to the equifinality problem often keep all possible parameter sets that could represent the system under study. The reason for this dissimilarity is a difference in study goal. From an optimization point of view, the solutions within our multiobjective ε-precision-defined ranges can be considered equal in the calibration objective space for the calibration period used. There is, therefore, no significant performance trade-off present and we can collapse them to a single solution. We do not claim that the actual parameter sets underlying these solutions are identical, or that these parameter sets might not produce greater performance differences for data periods that differ in their climatic regime. If our goal was to identify those parameter sets that are possible representations of the watershed, then this approach is not applicable. However, since we are only concerned with the level of trade-off present, we believe that our strategy provides a conceptually and computationally simple metric for assessment. Our approach is therefore not inconsistent with strategies like the limits of acceptability method [*Blazkova and Beven*, 2009; *Dean et al.*, 2009], or related approaches that we have promoted ourselves [*Yadav et al.*, 2007; *Zhang et al.*, 2008] to identify all suitable parameter sets.

[48] Some readers might also find the choice of ε-precision values to be somewhat arbitrary. The values we chose were based on rough estimates that would remain consistent with a relatively low level of error in the streamflow observations. However, these ε-precision values should probably be significantly larger if they were based on assumed errors in the observations of model forcing and response [*Liu et al.*, 2009; *Westerberg et al.*, 2010; *Krueger et al.*, 2009; *Kuczera et al.*, 2010; *Vrugt et al.*, 2008]. Our values are therefore very conservative estimates, which makes our conclusions all the more significant showing that even for small ε-precision values, we find that significant trade-offs do not exists for many watersheds.

[49] **Acknowledgments.** The authors of this work were partially supported by the U.S. National Science Foundation under grant EAR-0838357. Thorsten Wagener was partially funded by EPA STAR program grant RD834196. The computational resources for this work were provided in part through instrumentation funded by the National Science Foundation through grant OCI-0821527. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the U.S. National Science Foundation. The authors would also like to thank Keith Beven, Minxue He, Jasper Vrugt, and one anonymous reviewer for their helpful feedback in improving the quality of this manuscript.

References

Bekele, E. G., and J. W. Nicklow (2007), Multi-objective automatic calibration of SWAT using NSGA-II, *J. Hydrol.*, *341*(3–4), 165–176.

Bergström, S. (1975), The development of a snow routine for the HBV-2 model, *Nord. Hydrol.*, *6*(2), 73–92.

Bergström, S. (1992), The HBV model—Its structure and applications, *Rep. SMHI 4*, 35 pp., Swed. Meteorol. and Hydrol. Inst., Norrköping, Sweden.

Bergström, S. (1995), The HBV model, in *Computer Models of Watershed Hydrology*, edited by V. P. Singh, pp. 443–476, Water Resour. Publ., Highlands Ranch, Colo.

Beven, K. (2006), A manifesto for the equifinality thesis, *J. Hydrol.*, *320*, 18–36.

Beven, K., P. Smith, and A. Wood (2011), On the colour and spin of epistemic error (and what we might do about it), *Hydrol. Earth Syst. Sci. Discuss.*, *8*, 5355–5386.

Blazkova, S., and K. Beven (2009), A limits of acceptability approach to model evaluation and uncertainty estimation in flood frequency estimation by continuous simulation: Skalka catchment, Czech Republic, *Water Resour. Res.*, *45*(12), W00B16, doi:10.1029/2007WR006726.

Box, G. E. P., and D. R. Cox (1964), An analysis of transformations, *J. R. Stat. Soc. Ser. B*, *26*(2), 211–252.

Boyle, D., H. Gupta, and S. Sorooshian (2000), Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods, *Water Resour. Res.*, *36*, 3663–3674.

Boyle, D. P., H. V. Gupta, and S. Sorooshian (2003), Multicriteria calibration of hydrologic models, in *Calibration of Watershed Models, Water Sci. Appl.*, vol. 6, edited by Q. Duan et al., pp. 185–196, AGU, Washington, D. C., doi:10.1029/WS006p0185.

Cheng, C. T., C. P. Ou, and K. W. Chau (2002), Combining a fuzzy optimal model with a genetic algorithm to solve multi-objective rainfall-runoff model calibration, *J. Hydrol.*, *268*(1–4), 72–86.

Clark, M. P., A. G. Slater, D. E. Rupp, R. A. 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.

Coello Coello, C., G. B. Lamont, and D. A. Van Veldhuizen (2007), *Evolutionary Algorithms for Solving Multi-Objective Problems*, 2nd ed., 576 pp., Springer, N. Y.

Das, I. (1999), A preference ordering among various Pareto optimal alternatives, *Struct. Optimization*, *18*, 30–35.

Dean, S., J. Freer, K. Beven, A. Wade, and D. Butterfield (2009), Uncertainty assessment of a process-based integrated catchment model of phosphorus, *Stoch. Environ. Res. Risk Assess.*, *23*(7), 991–1010.

Deb, K. (2001), *Multi-Objective Optimization using Evolutionary Algorithms*, John Wiley, N. Y.

Deb, K., and R. Agrawal (1995), Simulated binary crossover for continuous search space, *Complex Systems*, *9*, 115–148.

de Vos, N. J., and T. H. M. Rientjes (2008), Multiobjective training of artificial neural networks for rainfall-runoff modeling, *Water Resour. Res.*, *44*(8), W08434, doi:10.1029/2007WR006734.

de Vos, N. J., T. H. M. Rientjes, and H. V. Gupta (2010), Diagnostic evaluation of conceptual rainfall-runoff models using temporal clustering, *Hydrol. Processes*, *24*(20), 2840–2850.

di Pierro, F. (2006), Many-objective evolutionary algorithms and applications to water resources engineering, Ph.D. thesis, School of Engineering, Computer Science and Mathematics, University of Exeter, Devon, U. K.

di Pierro, F., S. T. Khu, and D. Savic (2007), An investigation on preference order ranking scheme for multiobjective evolutionary optimization, *IEEE Trans. Evolut. Comput.*, *11*(1), 17–45.

Duan, Q., V. K. Gupta, and S. Sorooshian (1992), Effective and efficient global optimization for conceptual rainfall-runoff models, *Water Resour. Res.*, *28*(4), 1015–1031.

Duan, Q., et al. (2006), Model parameter estimation experiment (MOPEX): An overview of science strategy and major results from the second and third workshops, *J. Hydrol.*, *320*(1–2), 3–17.

Efstratiadis, A., and D. Koutsoyiannis (2010), One decade of multi-objective calibration approaches in hydrological modelling: A review, *Hydrol. Sci. J.*, *55*(1), 58–78.

Emsellem, Y., and G. de Marsily (1971), An automatic solution for the inverse problem, *Water Resour. Res.*, *7*(5), 1264–1283.

Farnsworth, R. K., and E. Thompson (1982), Mean monthly, seasonal, and annual pan evaporation for the United States, *NOAA Tech. Rep. NWS 34*, 88 pp., Natl. Weather Serv., Washington, D. C.

Fenicia, F., H. H. G. Savenije, P. Matgen, and L. Pfister (2007), A comparison of alternative multiobjective calibration strategies for hydrological modeling, *Water Resour. Res.*, *43*(3), W03434, doi:10.1029/2006WR005098.

Gill, M. K., Y. H. Kaheil, A. Khalil, M. McKee, and L. Bastidas (2006), Multiobjective particle swarm optimization for parameter estimation in hydrology, *Water Resour. Res.*, *42*(7), W07417, doi:10.1029/2005WR004528.

Goldberg, D. E. (2002), *The Design of Innovation: Lessons From and for Competent Genetic Algorithms*, 248 pp., Kluwer Acad., Norwell, Mass.

Gupta, H., S. Sorooshian, and P. Yapo (1998), Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information, *Water Resour. Res.*, *34*, 751–763.

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

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

Hamilton, A. S., D. G. Hutchinson, and R. D. Moore (2000), Estimating winter streamflow using conceptual streamflow model, *J. Cold Reg. Eng.*, *14*(4), 158–175.

Hamon, W. (1961), Estimating potential evopotranspiration, *J. Hydraul.*, *87*(HY3), 107–120.

Hanne, T. (1999), On the convergence of multiobjective evolutionary algorithms, *Eur. J. Oper. Res.*, *117*, 553–564.

Harlin, J., and C. S. Kung (1992), Parameter uncertainty and simulation of design floods in sweden, *J. Hydrol.*, *137*(1–4), 209–230.

Hastie, T., R. Tibshirani, and J. Friedman (2009), *14.3.12 Hierarchical Clustering*, chap. 14, 2nd ed., pp. 520–528, Springer, N. Y.

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

Kavetski, D., and M. P. Clark (2011), Numerical troubles in conceptual hydrology: Approximations, absurdities and impact on hypothesis testing, *Hydrol. Processes*, *25*(4), 661–670.

Kavetski, D., and G. Kuczera (2007), Model smoothing strategies to remove microscale discontinuities and spurious secondary optima in objective functions in hydrological calibration, *Water Resour. Res.*, *43*(3), W03411, doi:10.1029/2006WR005195.

Kavetski, D., G. Kuczera, and S. W. Franks (2006), Calibration of conceptual hydrological models revisited: 1. Overcoming numerical artefacts, *J. Hydrol.*, *320*(1–2), 173–186.

Khu, S. T., and H. Madsen (2005), Multiobjective calibration with Pareto preference ordering: An application to rainfall-runoff model calibration, *Water Resour. Res.*, *41*, W03004, doi:10.1029/2004WR003041.

Khu, S. T., H. Madsen, and F. di Pierro (2008), Incorporating multiple observations for distributed hydrologic model calibration: An approach using a multi-objective evolutionary algorithm and clustering, *Adv. Water Resour.*, *31*(10), 1387–1398.

Kokkonen, T., H. Koivusalo, A. J. Jakeman, and J. Norton (2006), Construction of a degree-day snow model in the light of the 10 iterative steps in model development, in Proceedings of the iEMSs Third Biennial Meeting: *"Summit on Environmental Modelling and Software"*, [CD-ROM], edited by A. Voinov, A. J. Jakeman, and A. E. Rizzoli, 12 pp., Int. Environ. Modell. and Software Soc., Burlington, VT.

Kollat, J. B., and P. Reed (2005), The value of online adaptive search: A comparison of NSGA-II, ε-NSGAII, and ε MOEA, in *Evolutionary Multi Criterion Optimization: Third International Conference (EMO 2005)*, edited by C. Coello Coello, A. Hernandez, and E. Zitzler, pp. 386–398, Lecture Notes in Computer Science Springer, Berlin, Germany.

Kollat, J. B., and P. M. Reed (2006), Comparing state-of-the-art evolutionary multi-objective algorithms for long-term groundwater monitoring design, *Adv. Water Resour.*, *29*(6), 792–807.

Kollat, J. B., and P. M. Reed (2007), A computational scaling analysis of multiobjective evolutionary algorithms in long-term groundwater monitoring applications, *Adv. Water Resour.*, *30*(3), 408–419.

Krueger, T., J. Quinton, J. Freer, C. Macleod, G. Bilotta, R. Brazier, P. Butler, and P. Haygarth (2009), Uncertainties in data and models to describe event dynamics of agricultural sediment and phosphorus transfer, *J. Environ. Qual.*, *38*(3), 1137–1148.

Kuczera, G., B. Renard, M. Thyer, and D. Kavetski (2010), There are no hydrological monsters, just models and observations with large uncertainties!, *Hydrol. Sci. J.*, *55*(6), 980–991.

Laumanns, M., L. Thiele, K. Deb, and E. Zitzler (2002), Combining convergence and diversity in evolutionary multiobjective optimization, *Evolut. Comput.*, *10*(3), 263–282.

Lawrence, D., I. Haddeland, and E. Langsholt (2009), Calibration of HBV hydrological models using pest parameter estimation, *Tech. Rep. 1-2009*, 45 pp., Norw. Water Resour. and Energy Dir., Oslo.

Liden, R., and J. Harlin (2000), Analysis of conceptual rainfall-runoff modelling performance in different climates, *J. Hydrol.*, *238*(3–4), 231–247.

Lindström G., J. Rosberg, and B. Arheimer (2005), Parameter precision in the HBV-NP model and impacts on nitrogen scenario simulations in the Rönneå River, Southern Sweden, *Ambio*, *34*(7), 533–537.

Liu, Y., J. Freer, K. Beven, and P. Matgen (2009), Towards a limits of acceptability approach to the calibration of hydrological models: Extending observation error, *J. Hydrol.*, *367*(12), 93–103.

Madsen, H. (2000), Automatic calibration of a conceptual rainfall-runoff model using multiple objectives, *J. Hydrol.*, *235*, 276–288.

Madsen, H., G. Wilson, and H. Ammentorp (2002), Comparison of different automated strategies for calibration of rainfall-runoff models, *J. Hydrol.*, *261*, 48–59.

McIntyre, N., H. Lee, H. Wheater, A. Young, and T. Wagener (2005), Ensemble predictions of runoff in ungauged catchments, *Water Resour. Res.*, *41*(12), W12434, doi:10.1029/2005WR004289.

Misirli, F., H. V. Gupta, S. Sorooshian, and M. Thiemann (2003), *Bayesian Recursive Estimation of Parameter and Output Uncertainty for Watershed Models, Water Science Series*, vol. 6, pp. 113–124, AGU, Washington, D. C.

Moore, R. J. (1985), The probability-distributed principle and runoff prediction at point and basin scales, *Hydrol. Sci. Bull.*, *30*(2), 273–297.

Moore, R. J. (2007), The PDM rainfall-runoff model, *Hydrol. Earth Syst. Sci.*, *11*(1), 483–499.

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

National Weather Service (2011), Model parameter estimation experiment (MOPEX), available at http://www.nws.noaa.gov/oh/mopex/mo_datasets.htm, Hydrologic Science and Modelling Branch, NWS Office of Hydrologic Development, Silver Springs, MD.

Neuman, S. (1973), Calibration of distributed parameter groundwater flow models viewed as a multiple-objective decision process under uncertainty, *Water Resour. Res.*, *9*(4), 1006–1021.

Nicklow, J. W., et al. (2010), State of the art for genetic algorithms and beyond in water resources planning and management, *J. Water Resour. Plan. Manage.*, *136*(4), 412–432.

Ogden, F. L., N. A. Abebe, and N. R. Pradhan (2010), Sensitivity and uncertainty analysis of the conceptual HBV rainfall-runoff model: Implications for parameter estimation, *J. Hydrol.*, *389*(3–4), 301–310.

Pareto, V. (1896a), *Cours D'Economie Politique*, vol. 1, 430 pp., Rouge, Lausanne, France.

Pareto, V. (1896b), *Cours D'Economie Politique*, vol. 2, 426 pp., Rouge, Lausanne, France.

Reed, P., J. B. Kollat, and V. Devireddy (2007), Using interactive archives in evolutionary multiobjective optimization: A case study for long-term groundwater monitoring design, *Environ. Model. Software*, *22*(5), 683–692.

Sawicz, K., T. Wagener, M. Sivapalan, P. A. Troch, and G. Carrillo (2011), Catchment classification: Empirical analysis of hydrologic similarity based on catchment function in the eastern USA, *Hydrol. Earth Syst. Sci. Discuss.*, *8*, 4495–4534.

Seibert, J. (1997), Estimation of parameter uncertainty in the HBV model, *Nord. Hydrol.*, *28*(4–5), 247–262.

Seibert, J. (2000), Multi-criteria calibration of a conceptual runoff model using a genetic algorithm, *Hydrol. Earth Syst. Sci.*, *4*(2), 215–224.

Seibert, J., and J. J. McDonnell (2002), On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria model calibration, *Water Resour. Res.*, *38*(11), 1241, doi:10.1029/2001WR000978.

Singh, S. (2010), Robust parameter estimation in gauged and ungaugedbasins, Ph.D. thesis, Institute for Modelling Hydraulic and Environmental Systems, University of Stuttgart, Germany.

Springer, S., J. Kite, and V. Schmidt (1997), Cave sedimentation, genesis, and erosional history in the Cheat River Canyon, West Virginia, *Geological Society of America Bulletin*, *109*(5), 524–532.

Sun, N.-Z. (1994), *Inverse Problems in Groundwater Modeling, Theory and Applications of Transparent Porous Media*, vol. 6, 352 pp., Kluwer Acad., N. Y.

Tang, Y., P. Reed, and T. Wagener (2006), How efficient and effective are evolutionary multiobjective algorithms at hydrologic model calibration?, *Hydrol. Earth Syst. Sci.*, *10*, 289–307.

Tang, Y., P. Reed, and J. B. Kollat (2007), Parallelization strategies for rapid and robust evolutionary multiobjective optimization in water resources applications, *Adv. Water Resour.*, *30*(3), 335–353.

Teytaud, O. (2007), On the hardness of offline multi-objective optimization, *Evolut. Comput.*, *15*(4), 475–491.

U.S. Geological Survey (USGS) (2010), Water-Data Report 2010 for 03274000 Great Miami River at Hamilton, OH, *Tech. Rep.*, 3 pp., USGS.

van Griensven, A., and W. Bauwens (2003), Multiobjective autocalibration for semidistributed water quality models, *Water Resour. Res.*, *39*(12), 1348, doi:10.1029/2003WR002284.

van Werkhoven, K., T. Wagener, P. Reed, and Y. Tang (2008), Characterization of watershed model behavior across a hydroclimatic gradient, *Water Resour. Res.*, *44*, W01429, doi:10.1029/2007WR006271.

van Werkhoven, K., T. Wagener, P. Reed, and Y. Tang (2009), Senitivity-guided reduction of parametric dimensionality for multi-objective calibration of watershed models, *Adv. Water Resour.*, *32*, 1154–1169.

Vorosmarty, C. J., C. A. Federer, and A. L. Schloss (1998), Evaporation functions compared on US watersheds: Possible implications for global-scale water balance and terrestrial ecosystem modeling, *J. Hydrol.*, *207*(3–4), 147–169.

Vrugt, J., H. V. Gupta, L. A. Bastidas, W. Bouten, and S. Sorooshian (2003), Effective and efficient algorithm for multiobjective optimization of hydrologic models, *Water Resour. Res.*, *39*(8), 1214, doi:10.1029/2002WR001746.

Vrugt, J., C. ter Braak, M. Clark, J. Hyman, and B. Robinson (2008), Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation, *Water Resour. Res.*, *44*(12), W00B09, doi:10.1029/2007WR006720.

Wagener, T., D. Boyle, M. J. Lees, H. S. Wheater, H. V. Gupta, and S. Sorooshian (2001), A framework for development and application of hydrological models, *Hydrol. Earth Syst. Sci.*, *5*(1), 13–26.

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

Wagener, T., H. S. Wheater, and H. V. Gupta (2004), *Rainfall-Runoff Modeling in Gauged and Ungauged Catchments*, 300 pp., Imperial College Press, London, U. K.

Wang, D., and M. Hejazi (2011), Quantifying the relative contribution of the climate and direct human impacts on mean annual streamflow in the contiguous United States, *Water Resour. Res.*, *47*(9), W00J12, doi:10.1029/2010WR010283.

Westerberg, I., J. Guerrero, J. Seibert, K. Beven, and S. Halldin (2010), Stage-discharge uncertainty derived with a non-stationary rating curve in the Choluteca River, Honduras, *Hydrol. Processes*, *25*(4), 603–613.

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*(8), 1756–1774.

Yapo, P. O., H. V. Gupta, and S. Sorooshian (1998), Multi-objective global optimization for hydrologic models, *J. Hydrol.*, *204*, 83–97.

Yilmaz, K. K., H. V. Gupta, and T. Wagener (2008), A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, *Water Resour. Res.*, *44*, W09417, doi:10.1029/2007WR006716.

Zhang, X. N., and G. Lindstrom (1996), A comparative study of a Swedish and a Chinese hydrological model, *Water Resour. Bull.*, *32*(5), 985–994.

Zhang, Z., T. Wagener, P. Reed, and R. Bhushan (2008), Reducing uncertainty in predictions in ungauged basins by combining hydrologic indices regionalization and multiobjective optimization, *Water Resour. Res.*, *44*, W00B04, doi:10.1029/2008WR006833.

J. B. Kollat, P. M. Reed, and T. Wagener, Department of Civil and Environmental Engineering, Pennsylvania State University, 212 Sackett Bldg., University Park, PA 16802, USA. (juk124@psu.edu)