Parameter regionalization of a monthly water balance model for the conterminous United States

A.R. Bock[1], L.E. Hay[2], G.J. McCabe[2], S.L. Markstrom[2], and R.D. Atkinson[3]

[1] U.S. Geological Survey, Colorado Water Science Center, Denver Federal Center, P.O. Box 25046, MS 415, Denver, Colorado, 80225, USA
[2] U.S. Geological Survey, National Research Program, Denver Federal Center, P.O. Box 25046, MS 413,Denver, Colorado, 80225, USA
[3] U.S. Environmental Protection Agency, Office of Water (4503-T), 1200 Pennsylvania Ave., Washington, DC, 20004, USA

Correspondence to: A. Bock (abock@usgs.gov)

Abstract

A parameter regionalization scheme to transfer parameter values from gaged to ungaged areas for a monthly water balance model (MWBM) was developed and tested for the conterminous United States (CONUS). The Fourier Amplitude Sensitivity Test, a global-sensitivity algorithm, was implemented on a MWBM to generate parameter sensitivities on a set of 109,951 hydrologic response units (HRUs) across the CONUS. The HRUs were grouped into 110 calibration regions based on similar parameter sensitivities. Subsequently, measured runoff from 1,575 streamgages within the calibration regions were used to calibrate the MWBM parameters to produce parameter sets for each calibration region. Measured and simulated runoff at the 1,575 streamgages showed good correspondence for the majority of the CONUS, with a median computed Nash-Sutcliffe Efficiency coefficient of 0.76 over all streamgages. These methods maximize the use of available runoff information, resulting in a calibrated CONUS-wide application of the MWBM suitable for providing estimates of water availability at the HRU resolution for both gaged and ungaged areas of the CONUS.

1 Introduction

The WaterSMART program (http://water.usgs.gov/watercensus/WaterSMART.html) was started by the United States (U.S.) Department of the Interior in February 2010. Under WaterSMART, the National Water Census (NWC) was proposed as one of the U.S. Geological Survey's (USGS) key research directions with a focus on developing new hydrologic tools and assessments. One of the major components of the NWC is to provide estimates of water availability at a sub-watershed resolution nationally (http://water.usgs.gov/watercensus/streamflow.html) with the goal of determining: (1) if the Nation has enough freshwater to meet both human and ecological needs and (2) if this water will be available to meet future needs. Streamflow measurements do not provide direct observations of water availability at every location of interest; approximately 72 percent (%) of land within the conterminous U.S. is gaged, with approximately 13% of these gaged areas being unaffected by anthropogenic effects (Kiang et al., 2013). This creates the challenge of determining the best method to transfer information from gaged catchments to data-poor areas where results cannot be calibrated or evaluated with measured streamflow (Vogel, 2006). This transfer of model parameter information from gaged to ungaged catchments is known as hydrologic regionalization (Bloschl and Sivapalan, 1995).

Many hydrologic regionalization methods have focused on developing measures of similarity between gaged and ungaged catchments using spatial proximity and physical characteristics. These methods are highly dependent on the complexity of the terrain and scale at which the relations are derived. Spatial proximity is considered the primary explanatory variable for hydrologic similarity (Sawicz et al., 2011) because of the first-order effects of climatic and topographic controls on hydrologic response. Close proximity, however, does not always result in hydrologic similarity (Vandewiele and Elias, 1995; Smakhtin, 2001; Ali et al., 2012).

Physical characteristics have been used as exploratory variables to develop a better understanding of the relation between model parameters that represent model function, and physical properties of the catchment (Merz and Bloschl, 2004). The relation between model parameters and the relevant physical characteristics, expressed for example as a form of multivariate regression, can be transferred to ungaged catchments (Merz and Bloschl, 2004). Model parameter definitions are by nature ambiguous and often difficult to correlate to a small number of meaningful variables such as physical and climatic characteristics (Zhang et al., 2008); some studies have found no significant correlation between catchment attributes and model parameters (Seibert, 1999; Peel et al., 2000), whereas others found that high correlation does not guarantee parameters that result in reliable model simulations of measured data (Sefton and Howarth, 1998; Kokkonen et al., 2003; Oudin et al., 2010). Physical and hydrologic characteristics also are used to derive measures of similarity (or dissimilarity) from multi-dimensional attribute space, which can be used to identify donor catchments (Qamar et al., 2015), or classify catchments into discrete regions or clusters (Oudin et al., 2008, 2010; Samuel et al., 2011). While these methods have indicated some success in simulating behavior of specific hydrologic components, such as base flow (Santhi et al., 2008) or monthly flow regimes (Qamar et al., 2015), other efforts utilizing discrete clusters performed poorly in explaining variability of measured streamflow (McManamay et al., 2011).

Two important components of the transfer of parameters to ungaged catchments are the identification of (1) influential (and non-influential) parameters, and (2) geographic extents and scales at which parameters exert control on model function. Reducing the number of parameters is important for calibration efficiency by reducing the structural bias of the model and the uncertainty of results where they cannot be verified or confirmed (Van Griensven et al., 2006). A high number of calibrated, poorly constrained parameters can often mask data or structural errors, which can go undetected and reduce the skill of the model in replicating results outside of calibration conditions (Kirchner, 2006; Bloschl et al., 2013; Bárdossy et al., 2015). This increases the potential for equifinality of parameter sets and higher model uncertainty that can be propagated to model results (Troch et al., 2003).

Sensitivity analysis (SA) has advanced the understanding of parameter influence on model behavior and structural uncertainty. SA measures the response of model output to variability in model input and/or model parameter values. SA partitions the total variability in the model response to each individual model parameter (Reusser et al., 2011) and results in a more-defined set of parameters and parameter ranges. Identification of sensitive parameters and their ranges is important for hydrologic model applications as key model parameters can vary spatially across physiographic regions, and also temporally (Tang et al., 2007; Guse et al., 2013).

Until recently, the high computational demands of SA have limited most implementations of hydrologic model SA to local sensitivity algorithms that evaluate a single parameter at a time (Tang et al., 2007). Global SA uses random or systematic sampling designs of the entire parameter space to quantify variation in model output (Van Griensven et al. 2006, Reusser et al. 2011). Some of these methods can account for parameter interaction and quantify sensitivity in non-linear systems. Global SA methods are computationally intensive (Cuo et al., 2011), but ever increasing computational efficiency has allowed for the development and application of a large number of global SA algorithms.

Previous work has suggested that isolating the key parameters that control model performance can be used to infer dominant physical processes in the catchment, as well as which components of the model dominate hydrologic response (Van Griensven et al. 2006, Tang et al., 2007, Reusser et al., 2011). To date, there has been little analysis of the use of SA for deriving measures of hydrologic similarity across catchments that can be applied towards hydrologic regionalization of model parameters. The spatially-distributed application of SA could be used to provide additional information for the delineation of homogeneous regions for parameter transfer based on similarity of model results from the SA. This strategy allows for the use of the existing model information and configuration to develop a calibration and regionalization framework without significantly changing the model structure or implementation

In this study, we present a hydrologic regionalization methodology for the CONUS that derived regions of hydrologic similarity based on the response of a Monthly Water Balance Model (MWBM) to parameter SA. Groups of streamgages within each region are calibrated together to define a single parameter set for each region. By extending model calibration to a large number of sites grouped by similarity through a quantified measure of model behavior, a more specific and constrained parameter space that fits each region can be identified.

2 Methods

2.1 Monthly Water Balance Model

The MWBM (Fig. 1) is a modular accounting system that provides monthly estimates of components of the hydrologic cycle by using concepts of water supply and demand (Wolock and

McCabe, 1999; McCabe and Markstrom, 2007). Monthly temperature (T) is used to compute potential evapotranspiration (PET) and to partition monthly precipitation (P) into rain and snow (Fig. 1). Precipitation that occurs as snow is accumulated in a snow pack (snow storage as snow water equivalent, or SWE); rainfall is used to compute direct runoff (R_{direct}) or overland flow, actual evapotranspiration (AET), soil-moisture storage recharge, and surplus water, which eventually becomes runoff (R) (Fig. 1). When rainfall for a month is less than PET, AET is equal to the sum of rainfall, snowmelt, and the amount of moisture that can be removed from the soil. The fraction of soil-moisture storage that can be removed as AET decreases linearly with decreasing soil-moisture storage; that is, water becomes more difficult to remove from the soil as the soil becomes drier and less moisture is available for AET. When rainfall (and snowmelt) exceeds PET in a given month, AET is equal to PET; water in excess of PET replenishes soil-moisture storage. When soil-moisture storage reaches capacity during a given month, the excess water becomes surplus and a fraction of the surplus ($R_{surplus}$) becomes R, while the remainder of the surplus is temporarily held in storage. The MWBM has been previously used to examine variability in runoff over the CONUS (Wolock and McCabe, 1999; Hay and McCabe, 2002; McCabe and Wolock, 2011a) and the global extent (McCabe and Wolock, 2011b). Table 1 lists the MWBM parameters, with definitions and parameter ranges for calibration.

The *Ppt_adj* and *Tav_adj* parameters specify seasonal adjustments for precipitation and temperature, respectively. The seasonal adjustment parameters were included to account for errors in the precipitation and temperature data used in this analysis. Sources of systematic and non-systematic errors of climate forcing data are well documented from the precipitation gage-derived sources (Groisman and Legates, 1994; Adam and Lettenmaier, 2003). Interpolation of these systematic errors from point-scale to gridded domains may propagate these biases, especially in complex terrain (Clark and Slater, 2006; Oyler et al., 2015). The use of adjustment factors allows uncertainty associated with forcing data and model parameter values to be treated separately (Vrught et al., 2008).

Figure 1. Conceptual diagram of the Monthly Water Balance Model (McCabe and Markstrom 2007). Processes influenced by model parameters used in Fourier Amplitude Sensitivity Test (FAST) are identified by green arrow and numbered (Table 1).

Table 1. Monthly Water Balance Model parameters and ranges.

The MWBM was applied to the CONUS with 109,951 hydrologic response units (HRUs) from the Geospatial Fabric (Viger and Bock, 2014), a national database of hydrologic features for national hydrologic modeling applications (Fig. 2). This HRU derivation is based on an aggregation of the NHDPlus dataset (USEPA and USGS, 2010), an integrated suite of geospatial data that incorporates features from the National Hydrography Dataset (http://nhd.usgs.gov/), the National Elevation Dataset (http://ned.usgs.gov/), and the Watershed Boundary Dataset (http://nhd.usgs.gov/wbd.html). The sizes of the HRUs range from less than 1 square kilometer (km^2) up to 67,991 km^2, with an average size of 74 km^2.

Inputs to the MWBM by HRU are: (1) monthly P (millimeters), monthly mean T (degrees Celsius), (2) latitude of the site (decimal degrees), (3) soil moisture storage capacity (millimeters), and (4) monthly coefficients for the computation of PET (dimensionless). Monthly P and mean T were derived from the daily time step, $1/8^{o}$ gridded meteorological data for the period of record from January 1949 through December 2011 (Maurer et al., 2002). Monthly P and T data were aggregated for each HRU using the USGS Geo Data Portal (*http://cida.usgs.gov/climate/gdp/*) (Blodgett et al., 2011). Latitude was computed from the centroid of each HRU. Soil moisture storage capacity was calculated using a 1 km^2 grid derived from the Soils Data for the Conterminous United States (STATSGO) (Wolock, 1997). The monthly PET coefficients were calculated by calibrating the Hamon PET values to Farnsworth et al. (1982) mean monthly free-water surface evapotranspiration. McCabe et al. (2015) describes these PET coefficient calculations in detail.

Figure 2. Hydrologic Response Units of the Geospatial Fabric, differentiated by color, overlain by NHDPlus region boundaries (R01-R18).

2.2 Fourier Amplitude Sensitivity Test

A parameter SA for the CONUS was conducted for the MWBM using the Fourier Amplitude Sensitivity Test (FAST) to identify areas of hydrologic similarity. FAST is a variance-based global sensitivity algorithm that estimates the contribution to model output variance explained by each parameter (Cukier et al. 1973, 1975; Saltelli et al. 2000). Advantages of using FAST over other SA methods are that FAST can calculate sensitivities in non-linear systems, and is extremely computationally efficient. The seasonal adjustment factors were not incorporated in the FAST analysis. We viewed the seasonal adjustment factors as related more to the forcing data, and for this application only parameters associated with model structure were included (first five parameters in Table 1).

FAST transforms a model's multi-dimensional parameter space into a single dimension of mutually independent sine waves with varying frequencies for each parameter, while using the parameter ranges to define each wave's amplitude (Cukier et al., 1973, 1975; Reusser et al., 2011). This methodology creates an ensemble of parameter sets numbering from 1 to N, each of which is unique and non-correlated with the other sets. Parameter sets are derived using the corresponding y-values along each parameter's sine wave given a value on the x-axis. The model is executed for all parameter sets using identical climatic and geographic inputs for each simulation. The resulting series of model outputs are Fourier-transformed to a power spectrum of frequencies for each parameter. Parameter sensitivity is calculated as the sum of the powers of the output variance for each parameter, divided by the sum of the powers of all parameters (Total Variance). The parameter sensitivities are scaled so that the sensitivities for all parameters sum to 1. Thus, parameters that explain a large amount of variability in the model output have higher values of (i.e. closer to 1) parameter sensitivity values.

FAST was implemented with the MWBM using the 'fast' library in the statistical software R (Reusser, 2012; R Core Team, 2013). Parameter ranges used by FAST for generating wave amplitudes of parameter ensembles across the CONUS were based on table 1. The 'fast' R package pre-determines the minimal number of runs necessary to estimate the sensitivities for the given number of parameters (Cukier et al., 1973). For our application we generated an ensemble of 1000 parameter sets (as compared to the minimally suggested number of 71 estimated by 'fast') to have the capability to compare results of different sensitivity analysis methods. The computational efficiency of the MWBM allowed the parameter sets to be simulated quickly through parallel processing.

Many applications of SA in hydrologic modeling have evaluated parameter sensitivity for measured streamflow using performance-based measures such as bias, root mean squared error

(RMSE), and the Nash-Sutcliffe Efficiency (NSE) (Nash and Sutcliffe, 1970; Moriasi et al., 2007). In this study, parameter sensitivity is examined using two hydroclimatic indices that account for the magnitude and variability of both climatic input and model output: the (1) Runoff Ratio (RR), a ratio of simulated runoff to precipitation, and (2) Runoff Variability (RV) index, the standard deviation of simulated runoff to the standard deviation of precipitation (Sankarasubramanian and Vogel, 2003).

3 Parameter regionalization procedure

The MWBM parameter sensitivities from the FAST analysis using an ensemble of 1000 MWBM parameter sets were evaluated across the CONUS. The spatial patterns and magnitudes of parameter sensitivities then were used to organize the 109,951 HRUs across the CONUS into hydrologically similar regions for parameter regionalization through MWBM calibration. Potential streamgages were identified for use in two automated calibration procedures. The calibration procedures were used to produce an 'optimal' set of MWBM parameters for each calibration region. The following sections describe the parameter regionalization procedure in detail (Fig. 3).

Figure 3. Schematic flowchart of the parameter sensitivity analysis and regionalization method described in this paper (Section 3).

3.1 Parameter sensitivities

The relative sensitivities derived from the FAST analysis using the RR and RV indices at each of the 109,951 HRUs across the CONUS were scaled so that the five MWBM parameter sensitivities derived for each HRU summed to 100 (Fig. 4). RR (Fig. 4a) is most sensitive to the parameter *Drofac* in regions where MWBM runoff is not dominated by snowmelt and orographic precipitation, such as arid and sub-tropical areas of the CONUS. MWBM parameters that control snowpack accumulation and melt (*Meltcoef*, *Tsnow*, and *Train*) are more important to the RR in the extensive mountain ranges in the Western CONUS, and northerly latitudes around the Great Lakes and in the Eastern CONUS. The RR indicates the highest sensitivity to the *Rfactor* parameter in mountainous areas of the CONUS and areas of the West Coast, and moderate to high sensitivity in areas where the sensitivity of RR to *Drofac* is low. *Tsnow*, *Train*, and

Meltcoef all share similar patterns across the CONUS. The spatial variability of the sensitivity of RR to *Meltcoef* indicates different physical mechanisms controlling *Metlcoef* parameter influence on RR in different areas of the CONUS. In the Western CONUS, the sensitivity of RR to *Meltcoef* is greatest in mountainous areas that accumulate and hold snowpack through the late spring, such as the Rocky Mountains, Cascade, and Sierra Nevada mountain ranges. In the Eastern and Midwestern CONUS, the sensitivity of RR to *Meltcoef* is greatest for HRUs with more northerly latitudes.

Figure 4. Relative sensitivity of the (a) Rainfall Ratio (RR) and (b) Runoff Variability (RV) indices to Monthly Water Balance Model parameters.

The spatial patterns of sensitivities of RV to the five MWBM parameters (Fig. 4b) show both similarities and deviations from the patterns shown in the RR maps. For the central part of the CONUS, the relative sensitivity for the parameter *Drofac* is high for both indices, and low for the parameter *Rfactor* for both indices. *Meltcoef*, *Tsnow*, and *Train* share the same relations between higher sensitivity and higher elevation (primarily in the western part of the CONUS), and higher sensitivity and more northerly latitude (primarily in the eastern half of the CONUS) for both indices. However, *Drofac* and *Rfactor* show distinctly different patterns of relative sensitivities for the eastern part of the CONUS for RV as compared to RR. The other three parameters follow the same general spatial patterns for RV as compared to RR, but with greater fine-scale spatial variation and patchiness. The differences between the spatial distributions of the sensitivities between the two indices highlight that applying SA to different model outputs can generate different levels of sensitivities for each parameter. In addition, the choice of objective function or model output for which to measure parameter sensitivity is important, as parameter sensitivities will differ depending on whether a user evaluating measures of magnitude, the variability of distribution, or timing (Krause et al., 2005; Kapangaziwiri et al., 2012).

Figure 5 illustrates the variability of parameter sensitivities between NHDPlus regions 08 (Lower Mississippi) and 14 (Upper Colorado) (see Fig. 2) for the RV and RR indices. The Lower Mississippi and Upper Colorado NHDPlus regions have a similar number of HRUs (4,449 and 3,879, respectively) and cover a similar area (26,285 and 29,357 km^2, respectively). The Lower Mississippi region has homogenous topography, with humid, subtropical climate, while the

Upper Colorado region has highly variable topography, and thus highly variable climatic controls on hydrologic processes. For the Lower Mississippi region only one parameter dominates modeled RV variance (*Rfactor*, Fig. 5a) and modeled RR variance (*Drofac*, Fig. 5c). In contrast, for the Upper Colorado River region several parameters influence RV variability (*Drofac*, *Rfactor* and *Meltcoef*, Fig. 5b) and RR variability (*Drofac* and *Meltcoef*, Fig. 5d). In the Lower Mississippi Region, the amount of snowfall is negligible, so the three parameters that control snowfall and snowpack accumulation in the MWBM have negligible effect on simulated total runoff. The comparison of the parameter sensitivities for these two regions illustrates how variable parameter sensitivities are for different regions (i.e. different climatic and physiographic regions)

Figure 5. Parameter sensitivities of Runoff Variability (RV; a-b) and Runoff Ratio (RR; c-d) indices for Monthly Water Balance Model parameters in the Lower Mississippi (R08) and Upper Colorado (R14).

3.2. Calibration regions

The spatial patterns and magnitudes of parameter sensitivities across the CONUS were used as a basis for organizing HRUs into hydrologically similar regions for parameter regionalization through MWBM calibration. This idea is rooted in the hypothesis that geographically proximate HRUs share similar forcings and conditions, and thus will behave similarly. This application uses similarity in SA results as a basis for organization, rather than similarity in physiographic characteristics. The derived regions are subsequently used to simplify model calibration across the CONUS and provide a basis for the transfer and application of parameters to ungaged areas.

The parameter sensitivities derived from the RR were used to organize the HRUs into two independently-derived calibration regions; the first derived by identifying HRUs with unique combinations of the order of parameter sensitivities to the RR (highest parameter sensitivities to lowest, i.e. 1-*Drofac* (78%), 2-*Rfactor* (16%), 3-*Meltcoef* (4%), 4-*Tsnow* (1%), 5-*Train* (1%)), and the second classification based upon identifying HRUs with unique sets of parameters whose sensitivities exceeded a specified threshold of parameter sensitivity (i.e. only *Drofac* and *Rfactor* using a 5% threshold in the first classification example). This classification identified 16 distinct regions of HRUS across the CONUS based on the order of the parameter sensitivities of the five parameters (derived using the RR index). Sizes of these regions ranged from 94 km^2 to almost 2 million km^2. The second classification delineated regions with an identical set of the most important parameters to the RR based on parameters whose sensitivities exceeded a 5% threshold. This step identified 12 regions of HRUs with unique combinations of parameter sensitivities exceeding 5%. There has been progress in providing quantitative thresholds for the identification of sensitive and non-sensitive parameters for hydrologic modelers (Tang et al., 2007), but no definitive consensus yet exists. Therefore a 5% threshold was used based on visual delineation of major physiographic features such as mountain ranges across the CONUS. The sizes of this second group of regions ranged from 94 km^2 to more than 15 million km^2. Maps of the two groupings of HRUS were intersected to create a total of 49 regions across the CONUS. NHDPlus region boundaries, proximity, and significant topographic divides were used to further divide the groups into 159 geographically unique calibration regions across the CONUS. The lack of streamgages available in some regions, especially areas with arid and semi-arid climates, necessitated merging regions together. Calibration regions that contained less than 3 streamgages from the 8,410 gages present in the Geospatial Fabric (see section 3.3) were combined with the proximate and most similar group which shared the most similar parameter sensitivities (both order and magnitude), resulting in 110 calibration regions across the CONUS (Fig. 6). Within each region the FAST results for both the RR and RV indices were used to determine which parameters to calibrate. Parameters with a median parameter sensitivity of 5% for the RR and RV among the region's HRUs were selected for group calibration. Parameters not shown as sensitive were kept at the default value for the group.

Figure 6. Final 110 Monthly Water Balance Model calibration regions differentiated by colors. A subset of streamgages within each calibration region were calibrated in a group-wise fashion to produce a single optimized parameter set for the entire region (Fig. 3).

3.3 Initial streamgage selection

The initial set of streamgages used for testing in the MWBM calibration procedures was selected from 8,410 streamgages identified in the Geospatial Fabric (Fig. 7). The Geospatial Fabric includes reference and non-reference streamgages from the Geospatial Attributes of Gages for

Evaluating Streamflow dataset (GAGES-II, Falcone et al., 2010). Of the 8,410 streamgages in the Geospatial Fabric, 1,864 were identified as having reference-quality data with at least 20 years of record. These reference quality streamgages were judged to be largely free of human alterations to flow (Falcone et al., 2010). In the current study, reference quality was not considered in the initial streamgage selection because the 20 years of record was considered too restrictive. Therefore a subset of the 8,410 streamgages was selected for initial testing in the MWBM calibration procedures based on the following criteria:

(1) Remove streamgages with less than 10 years of total measured streamflow (120 months) within the time period 1950 – 2010.

(2) Remove streamgages with a drainage area defined by the Geospatial Fabric that are not within 5% of the USGS National Water Information System (NWIS) reported drainage area (U.S. Geological Survey, 2014). This eliminated many of the streamgages with smaller drainage areas due to the resolution of the Geospatial Fabric.

(3) Remove streamgages that did not have at least 75% of its drainage area contained within a single calibration region.

These criteria resulted in 5,457 potential streamgages for testing in the MWBM calibration procedures (Fig. 7). Streamflow at these streamgages was aggregated and converted from daily (cubic feet/second) to a monthly runoff depth (mm) (streamflow per unit area).

Figure 7. Streamgages tested in the study. GF notes geospatial fabric for national hydrologic modeling (Viger and Bock, 2014).

3.4 Monthly Water Balance Model calibration

Two automated calibration procedures were implemented to produce an ‘optimal’ set of MWBM parameters for each calibration region. The first procedure, Individual Streamgage Calibration, calibrated each of the 5,457 streamgages individually. Results from the individual calibrations were used to further filter the streamgages within the second procedure, Grouped Streamgage Calibration, which calibrated selected streamgages together by calibration region.

3.4.1 Individual streamgage calibration

The first calibration procedure was an automated process that individually calibrated each of the 5,457 streamgages from the initial streamgage selection with measured streamflow (U.S. Geological Survey, 2014). Results from these individual streamgage calibrations quantified the ‘best’ performance of the MWBM at each gage, providing a ‘baseline’ measure for evaluation.

The Shuffled Complex Evolution (SCE) global-search optimization algorithm (Duan et al., 1993) has been frequently used as an optimization algorithm in hydrologic studies (Hay et al., 2006; Blasone et al., 2007; Arnold et al., 2012), including previous studies with the MWBM (Hay and McCabe, 2010). Further details can be found in Duan et al. (1993). SCE was used to maximize a combined objective function based on: (1) Nash-Sutcliffe Efficiency (NSE) coefficient using measured and simulated monthly runoff and (2) NSE using natural log-transformed measured and simulated runoff (logNSE), using the entire period of record for each streamgage. The NSE measures the predictive power of the MWBM in matching the magnitude and variability of the measured and simulated runoff (Nash and Sutcliffe, 1970). The NSE coefficient ranges from $-\infty$ to 1, with 1 indicating a perfect fit, and values less than 0 indicating that measured mean runoff is a better predictor than model simulations. The NSE has been shown to give more weight to the larger values in a time series (peak flows) at the expense of lower values (low flows) (Legates and McCabe, 1999), so the logNSE was incorporated into the objective function to give weight to low-flow periods (Tekleab et al., 2011).

3.4.2 Grouped streamgage calibration

The second calibration procedure was an automated process that calibrated groups of streamgages together for each calibration region to derive a single set of MWBM parameters (Table 1) for each calibration region (Fig. 6). The NSE and logNSE values from the individual streamgage calibrations (described in the previous section) were used to identify streamgages that should not be used for grouped streamgage calibration. If the individual streamgage calibration was not ‘satisfactory’, then it was felt that it would not provide useful information for the grouped streamgage calibration procedure.

Satisfactory individual streamgage calibrations were identified with the following procedure:

(1) Eliminate all streamgages with NSE values < 0.3.

(2) If the number of remaining streamgages for a given calibration region is > 10, then eliminate all streamgages with NSE < 0.5.

(3) If the number of streamgages for a given calibration region is > 25, then eliminate all streamgages with NSElog < 0.

(4) If the number of remaining streamgages for a calibration region is < 5, check to see if any of the eliminated streamgages were reference streamgages (as defined in Falcone et al., 2010), then add the reference streamgages back in if the NSE value > 0.0. Reference streamgages are USGS streamgages deemed to be largely free of anthropogenic impacts and flow modifications (Falcone et al., 2010; Kiang et al., 2013).

These criteria, while somewhat arbitrary, were chosen so that no calibration region had less than 5 streamgages for the grouped streamgage calibration. Using the above criterion, of the 5,457 streamgages individually calibrated, 3,125 remained as candidates for the grouped streamgage calibration procedure.

The grouped streamgage calibration procedure used the SCE global-search optimization algorithm with a multi-term objective function (Eq. 1). Measured and simulated values for selected streamgages contained within a calibration region were scaled to Z-scores to remove differences in magnitudes between streamgages (Eq. 2). The multi-term objective function minimized the sum of the absolute differences between Z-scores from four measured and simulated time series: mean monthly runoff (MMO, MMS), monthly runoff (MO, MS), annual runoff (AO, AS) (U.S. Geological Survey, 2014), and monthly snow water equivalent (SO, SS) for all selected streamgages within a given calibration region:

$$min\sum_{i=1}^{n}[3|\mathrm{MMO_i}-\mathrm{MMS_i}|+|\mathrm{MO_i}-\mathrm{MS_i}|+|\mathrm{AO_i}-\mathrm{AS_i}|+0.5|\mathrm{SO_i}-\mathrm{SS_i}|] \qquad \text{(Eq.1)}$$

$$\text{where}\begin{cases} 0 \text{ if } 0.75 < \mathrm{SO_i}-\mathrm{SS_i} < 1.25 \\ |\mathrm{SO_i}-\mathrm{SS_i}| \; if \; \mathrm{SS_i} < \mathrm{SO_i^{0.75}} \\ |\mathrm{SO_i}-\mathrm{SS}_i| \; \mathrm{SS_i} > \mathrm{SO_i^{1.25}} \end{cases}$$

The measured and simulated Z-scores were calculated as:

$$Z = (x-u)/\sigma \qquad \text{(Eq. 2)}$$

where x is the time-series value, u is the mean, and σ the standard deviation of the measured and simulated variable.

'Measured' SWE was determined for each HRU from the Snow Data Assimilation System (SNODAS; National Operational Hydrologic Remote Sensing Center, 2004) and included a +/-25% error bound. The unconstrained automated calibration (without a restriction on SWE) led to unrealistic sources of snowmelt in the summer that enhanced the low-flow simulations. The 25% error bound is arbitrary; calibrating to the actual SNODAS SWE values was found to be too restrictive, but adding this error bound to the SWE values resulted in better overall runoff simulations. The absolute differences of the simulated SWE Z-scores within +/- 25% of the measured SWE Z-score were designated as 0. Otherwise, the absolute difference was computed between the simulated SWE Z-score and either the upper or lower bounds (Eq. 1).

The grouped calibration procedure was run for all 110 calibration regions. For each calibration region the seasonal adjustment parameters and the sensitive parameters (identified by the FAST analysis -- section 3.1) were calibrated; parameters deemed not sensitive (parameter sensitivity < 5% of total variance) were set to their default values (see Table 1). The entire period of the streamflow record for each streamgage was split by alternating years. After calibration, mean monthly measured and simulated Z-scores for runoff at all selected streamgages within a calibration region were compared.

Figure 8 shows an example of the graphic used to evaluate the measured and simulated mean monthly Z-scores for 21 streamgages selected for the region located in the Tennessee River calibration region (part of NHDPlus Region R06 in Fig. 2); the orange, red, and black dots indicate calibration, evaluation, and the entire period of record, respectively. A tight grouping around the one-to-one line indicates good correspondence between measured and simulated Z-scores. Points closer to the upper right corner of each plot represent high-flow periods. Points closer to the lower left corner of the plot represent low-flow periods. Streamgages within a calibration region were assigned the same parameter values; therefore streamgages that plotted outside (two standard deviations) of the one-to-one line were considered to not be representative of the calibration region, and the calibration procedure for that calibration region was repeated without those streamgages.

Figure 8. Measured versus simulated mean monthly Z-scores for the Tennessee River calibration region (see Fig. 10b for location). Orange is calibration, red is evaluation, and black is all years.

The goal of the second calibration procedure was to find a single parameter set for each calibration region. Past applications of the MWBM (Wolock and McCabe, 1999, McCabe and Wolock, 2011a) used a single set of fixed MWBM parameters for the entire CONUS. Many of the streamgages included in the second calibration procedure could be affected by significant anthropogenic effects; the seasonal adjustment factors, calibrated at each individual streamgage, could account for these effects and result in satisfactory NSE values. Streamgages that were removed due to poor performance in the second calibration were assumed to have anthropogenic effects not consistent with the streamgages that plotted along the one-to-one line. Poor performance may result because the MWBM fails to reliably simulate runoff for a watershed because of model limitations (i.e. not including all important hydrologic processes), but the calibration regions are assumed to be homogeneous based on the FAST analysis. Therefore it is assumed that if some of the streamgages within a region have satisfactory results, then the MWBM is able to simulate runoff in that region.

4 MWBM calibration region results

4.1 Individual streamgage calibration results

The individual streamgage calibrations provided information regarding: (1) the potential suitability of a given streamgage for inclusion in a grouped calibration, and (2) a ‘baseline’ measure for evaluation of the grouped calibration results. Reference and non-reference streamgages were considered in this application; if the runoff at a streamgage could not be calibrated individually to a ‘satisfactory’ level (based on criterion outlined in section 3.4.2), then it was felt that it would not provide useful information for the grouped streamgage calibration procedure. Figure 9 shows the NSE (Fig. 9a) and logNSE (Fig. 9b) coefficients from the individual streamgage calibrations for the CONUS. Scattered throughout the CONUS are NSE and logNSE values less than 0.0 (triangles in Fig. 9). These poor results are likely streamgages with poor streamflow records, either due to measurement error or anthropogenic effects (dams, water use, etc.).

Figure 9. Individual streamgage calibration results: (a) Nash-Sutcliffe Efficiency (NSE) coefficient and (b) log of the NSE (logNSE).

4.2 Grouped streamgage calibration results

4.2.1 Mean monthly z-scores

Figure 10a shows a scatterplot of measured versus simulated mean monthly Z-scores for runoff, similar to Figure 8, but based on all available years (the black dots in Fig. 8) for all the final calibration streamgages (1,575 streamgages). Four regions are highlighted to illustrate the monthly variability in MWBM results across the CONUS (see Fig. 10b for locations). The four regions are: New England (67 streamgages, red); Tennessee River basin (21 streamgages, orange); Platte Headwaters (15 streamgages, blue); and Pacific Northwest (33 streamgages, green) (Fig. 10b).

Figure 10. (a) Measured versus simulated mean monthly Z-scores for runoff at all streamgages and (b) location of highlighted streamgages for four calibration regions: New England (67 streamgages, red); Tennessee River (21 streamgages, orange); Platte Headwaters (15 streamgages, blue); and Pacific Northwest (33 streamgages, green).

In Fig. 10a, three of the regions (New England, Tennessee River, and Pacific Northwest), show simulated Z-scores that correspond favorably to measured Z-scores for each of the twelve months, including periods of low and high runoff. These regions represent marine or humid climates with homogenous physio-climatic conditions and an even spatial distribution of streamgages, where models should be expected to perform well (see Fig. 9) There is a higher variability in model results for the high-flow months (May - June) for streamgages within the

Platte Headwaters (Fig. 10a; blue dots) than for low-flow months. This variability may be related to factors controlling the magnitude and timing of snow melt runoff (Fig. 9).

For each calibration streamgage, a set of four months were identified that represent different parts of the measured mean monthly hydrograph (highest- and lowest- flow month and the two median-flow months). The measured and simulated mean monthly streamflow Z scores corresponding to the four months are plotted as cumulative frequencies (Fig. 11) to compare how well the simulated Z scores matched measured Z scores for different parts of the hydrograph over the entire set of calibration gages. For the highest-flow, there is an under-estimation of runoff, with the greatest divergence between the two distributions in the middle to lower half of the distribution (Fig. 11a). For the median-flow, the measured and simulated Z scores are well matched (Fig. 11b). For the 10 lowest-flow, simulated Z scores are greater than measured Z scores, with the greatest divergence between the two distributions in the middle to upper half of the distribution (Fig. 11c).

Figure 11. Z-score cumulative frequency for (a) highest-, (b) median-, and (c) lowest-flow months.

The median Z-score errors (simulated - measured) by region for the (a) highest-, (b) median-, and (c) lowest-flows are shown in Figure 12. The largest errors are for the highest-flows (Fig. 12a). The MWBM simulations under-estimate the highest flows for much of the CONUS. The errors for median-flows are fairly uniform and consistent across the CONUS (Fig. 12b), with a median error close to 0. For the lowest-flow months the MWBM over-estimates low flows for a large portion of the Midwest (Fig. 12c).

Figure 12. Z-score error (simulated - measured) for (a) highest-, (b) median-, and (c) lowest-flow months.

4.2.2 Nash-Sutcliffe efficiency

Figure 13 compares the NSE from the individual streamgage calibrations (gageNSE) with the grouped calibrations (groupNSE) for all final streamgages used in the second calibration procedure. NSE values > 0.75 (dashed line) and > 0.5 (solid line) indicate very good and satisfactory results (Moriasi et al., 2007). Overall, most NSE values fall above the 0.5 NSE threshold of satisfactory performance (median of gageNSE and groupNSE = 0.76). The gageNSE values are used here as a 'baseline' for evaluation of the groupNSE results. The groupNSE values were not expected to be greater than the gageNSE values since (1) NSE was not used as an objective function in the grouped calibration, and (2) grouped calibrations found the 'best' parameter set for a set of streamgages versus an individual streamgage. Figure 13 shows an equal distribution of NSE values around the one-to-one line, indicating that the grouped calibration provided additional information over the individual streamgage calibrations (cases where groupNSE are greater than gageNSE in Fig. 13). The difference between the gageNSE and groupNSE becomes larger as the NSE values decrease, reflecting the increasing uncertainty in the grouped calibrations in areas with lower gageNSE values.

Figure 13. Nash Sutcliffe Efficiency from individual (gageNSE) and grouped (groupNSE) calibration. Calibration regions in New England (67 streamgages, red); Tennessee River (21 streamgages, orange); Platte Headwaters (15 streamgages, blue); and Pacific Northwest (33 streamgages, green) are highlighted (see Fig. 10b for location).

Four regions are highlighted in Fig. 13 to illustrate the variability of NSE across the CONUS (see Fig. 10b for locations). The highlighted regions in New England (red), Tennessee River (orange), and Pacific Northwest (green), show good groupNSE and gageNSE results. Four of the 15 streamgages in the Platte Headwaters (blue) have groupNSE values $\leq$ 0.5. This is probably related to simulation error during the snowmelt period (May - June, Fig. 10a).

Figure 14 shows the median groupNSE by calibration region for the CONUS. The pattern is very similar to that shown for the individual streamgage calibration results in Fig. 9a and highlights the problem areas shown in Fig. 12.

Figure 14. Median Nash Sutcliffe Efficiency (NSE) by calibration region of streamgages used for calibration.

5 Discussion

This study presented a parameter regionalization procedure for calibration of the MWBM, resulting in an application that can be used for simulation of hydrologic variables for both gaged and ungaged areas in the CONUS. The regionalization procedure grouped HRUs on the basis of similar sensitivity to five model parameters. Parameter values and model uncertainty information within a group was then passed from gaged to ungaged areas within that group.

5.1 Regionalized parameters

Results from this study indicate that regionalized parameters can be used to produce satisfactory MWBM simulations in most parts of the CONUS (Fig. 13). Despite the differences between the individual streamgage calibration and grouped calibration, Figure 13 illustrates that the grouped calibration strategy, which focused only on sensitive parameters, can provide just as much information as the individual streamgage calibration with no constraints on the parameter optimization other than the default ranges. The MWBM is a simple hydrologic model as it has minimal parameters, which are conceptual in nature (not physically based). It may be that this type of model is best for regionalization when parameter sensitivity can be identified and HRU behavior can be classified by a small number of clearly defined spatial groups. More complicated models with many more interactive parameters may not respond as well to this simple type of regionalization; more parameters may lead to more parameter interaction and situations of equifinality which might confuse the analysis.

The adjustments of precipitation and temperature parameters for the individual streamgage calibrations accounted for local errors such as rain gage under catch of precipitation. In addition these climate adjustments also account for local anthropogenic effects on streamflow (e.g. dams, diversions) since streamgages were not screened for these effects prior to individual streamgage calibration. In the grouped streamgage calibrations, the same precipitation and temperature adjustments are applied at every streamgage within the calibration region, making these climate adjustments more of a regional adjustment and producing more of a ‘reference’ condition for each calibration region.

5.2 Parameter sensitivities and dominant process

The MWBM parameter sensitivities varied by hydroclimatic index (RR and RV) and across the CONUS (Fig. 4). The parameter sensitivity patterns give an indication of dominant hydrologic processes based on MWBM. The dominant process can be seasonal and MWBM performance may be enhanced by extending the use of SA along the temporal domain to identify and temporally vary the parameters that are seasonally important to the MWBM. For example, error in peak flow months is the primary cause for poor model performance in the Platte Headwaters (Fig. 10). For the Platte Headwaters, the final parameter set performed well for simulated Z-scores for the regionalized low- and median-flow conditions (Fig. 10a, July through April), but was not able to replicate measured mean monthly flows for May and June. In this case, the dominant processes controlling hydrologic behavior change with season and the parameters controlling the dominant response may have to change accordingly (Gupta et al., 2008; Reusser et al., 2011).

5.3 Model accuracy

The pattern of MWBM accuracies shown in Fig. 9 and 14 are similar to those shown by Newman et al. (2015; Fig. 5a) in which a daily time-step hydrologic model was calibrated for 671 basins across the CONUS. Our study and the Newman et al. (2015) study both indicate the same 'problem areas' with the poorest performing basins generally being located in the high plains and desert southwest. Newman et al. (2015) attributed variation in model performance by region to spatial variations in aridity and precipitation intermittency, contribution of snowmelt, and runoff seasonality.

The inferior MWBM results in the 'problem areas' can be attributed to multiple factors which likely include inadequate hydrologic process representation and errors in forcing data (e.g. climate data), and/or measured streamflow. Archfield et al. (2015) state that the performance of continental-domain hydrologic models is considerably constrained by inadequate model representation of dominant hydrologic processes. For example, the simplicity of the MWBM presents limitations on the representation of deeper groundwater reservoirs, gaining and losing stream reaches, simplistic AET, and the effects of surface processes (infiltration and overland flow) that need to be represented at finer time steps than monthly.

The dominant hydrologic processes in the ‘problem areas’ appear to be poorly represented at the daily (Newman et al., 2015) and monthly time steps. This may be due to inadequate forcing data, the quality of which ‘is paramount in hydrologic modeling efforts’ (Archfield et al., 2015) and/or the lack of ‘good’ reference streamflow data for calibration and evaluation. Both surely play a role and emphasize the need for incorporation of additional datasets so that calibration and evaluation of intermediate states in the hydrologic cycle are examined.

6 Conclusions

A parameter regionalization procedure was developed for the CONUS that transferred parameter values from gaged to ungaged areas for a MWBM. The FAST global-sensitivity algorithm was implemented on a MWBM to generate parameter sensitivities on a set of 109,951 HRUs across the CONUS. The parameter sensitivities were used to group the HRUs into 110 calibration regions. Streamgages within each calibration region were used to calibrate the MWBM parameters to produce a regionalized set of parameters for each calibration region. The regionalized MWBM parameter sets were used to simulate monthly runoff for the entire CONUS. Results from this study indicate that regionalized parameters can be used to produce satisfactory MWBM simulations in most parts of the CONUS.

The best MWBM results were achieved simulating low- and median-flows across the CONUS. The high-flow months generally showed lower skill levels than the low- and median-flow months, especially for regions with dominant seasonal cycles. The lowest MWBM skill levels were found in the high plains and desert southwest and can be attributed to multiple factors which likely include inadequate hydrologic process representation and errors in forcing data and/or measured streamflow. Calibration and evaluation of intermediary fluxes and states in the MWBM through additional measured datasets may help to improve MWBM representations of these model states by helping to constrain parameterization to measured values.

7 Acknowledgments

This research was financially supported by the U.S. Department of Interior South Central Climate Science Center (*http://southcentralclimate.org/*), U.S. Environmental Protection Agency Office of Water, and the U.S. Geological Survey WaterSMART initiative. This paper is a product of discussions and activities that took place at the USGS John Wesley Powell Center for Analysis and Synthesis (*https://powellcenter.usgs.gov/*). Further project support was provided by the Jeff Falgout of the USGS Core Science Systems (CSS) Mission Area. Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

8 References

Adam, J.C., and Lettenmaier, D.P.: Bias correction of global gridded precipitation for solid precipitation undercatch, J. Geophys. Res., 108, 1-14, doi:10.1029/2002JD002499, 2003.

Ali, G., Tetzlaff, D., Soulsby, C., McDonnell, L.L., and Capell, R.: A comparison of similarity indices for catchment classification using a cross-regional dataset, Adv. Water Resources, 40, 11-22, http://dx.doi.org/10.1016/j.advwatres.2012.01.008, 2012.

Archfield, S.A., Clark, M., Areheimer, B., Hay, L.E., McMilan, H., Kiang, J.E., Seibert, J., Bock, A., Wagener, T., Farmer, W.H., Andressian, V., Attinger, S., Viglione, A., Knight, R., Markstrom, S., and Over, T.: Improving the performance of hydrologic models across local to continental domains: A discussion of research needs, Water Resour. Res., 2015, In review.

Arnold, J.G., Moriasi, D.N., Gassman, P.W., Abbaspour, K.C., White, M.J., Srinivasan, R., Santhi, C., Harmel, R.D., van Griensven, A., Van Liew, M.W., Kannan, N., and Jha, M.K.: SWAT: Model Use, Calibration and Validation, T. ASABE, 55(4), 1491-1508, 2012.

Bárdossy, A., Huang, Y., and Wagener, T.: Simultaneous calibration of hydrological models in geographical space, Hydrol. Earth Syst. Sci. Discuss., 12, 1123-11268, doi:10.5194/hessd-12-11223-2015, 2015.

Blasone, R.-S., Madsen, H., and Rosbjerg, D.: Parameter estimation in distributed hydrological modelling: comparison of global and local optimisation techniques, Nord. Hydrol., 34,451-476, doi:10.2166/nh.2007.024, 2007.

Blodgett, D.L., Booth, N.L., Kunicki, T.C., Walker, J.L., and Viger, R.J.: Description and Testing of the Geo Data Portal: A Data Integration Framework and Web Processing Services for Environmental Science Collaboration. US Geological Survey, Open-File Report 2011-1157, 9 pp., Middleton, WI, USA, 2011.

Bloschl, G., and Sivapalan, M.: Scale issues in hydrological modeling: a review, Hydrol. Process., 9, 251-290, 1995.

Blosch, G., Sivapalan, M., Wagener, T., Viglione, A., and Savenije, H (Eds.): Runoff Prediction in Ungauged Basins: Synthesis across Processes, Places, and Scales. Cambridge University Press, Cambridge, England, 2013.

Clark, M.P., and Slater, A.G.: Probabilistic Quantitative Precipitation Estimation in Complex Terrain, B. Am. Meterol. Soc., 7, 3-2, doi: http://dx.doi.org/10.1175/JHM474.1, 2006.

Cukier, R.I., Fortuin, C.M., Shuler, K.E., Petschek, A.G, and Schaibly, J.H: Study of sensitivity of coupled reaction systems to uncertainties in rate coefficients 1, J. Chem. Phys., 59(8), 3873-3878, 1973.

Cukier, R.I., Schiably, J.H., and Shuler, K.E: Study of sensitivity of coupled reaction systems to uncertainties in rate coefficients 3, J. Chem. Phys., 63(3), 1140-1149, 1975.

Cuo, L., Giambelluca, T.W., and Ziegler, A.D: Lumped parameter sensitivity analysis of a distributed hydrological model within tropical and temperate catchments, Hydrol. Process., 25(15), 2405-2421, doi:10.1002/hyp.8017, 2011.

Duan, Q., Gupta, V.K., and Sorooshian, S.: A shuffled complex evolution approach for effective and efficient optimization, J. Optimiz. Theory App., 76, 501-521, doi:10.1007/BF00939380, 1993.

Falcone, J.A., Carlisle, D.M., Wolock, D.M., and Meador, M.R.:   GAGES: A stream gage database for evaluating natural and altered flow conditions in the conterminous United States, Ecology, 91, p. 621, A data paper in Ecological Archives E091-045-D1, available at http://esapubs.org/Archive/ecol/E091/045/metadata.htm (last accessed 15 November 2012), 2010.

Farnsworth, R.K., Thompson, E.S., and Peck, E.L.: Evaporation Atlas for the Contiguous 48 United States, NOAA Technical Report NWS 33, 41 pp., National Oceanic and Atmospheric Administration, Washington, D.C., 1982.

Groisman, P.Y., and Legates, D.R.,: The accuracy of United States precipitation data, Bull. Am. Meterol. Soc., 75(2), 215-227, doi:10.1029/1998JD200110, 1994.

Gupta, H.V., Wagener, T., Liu, Y.Q. "Reconciling theory with observations: Elements of diagnostic approach to model evaluation." Hydrologic Processes (2008): 22,3802-3813.

Guse, B., Reusser, D.E., and Fohrer, N.: How to improve the representation of hydrological processes in SWAT for a lowland catchment - temporal analysis of parameter sensitivity and model performance, Hydrol. Process., 28(4), 2561-2670, doi:10.1002/hyp.9777, 2013.

Hay, L.E., Leavesley, G.H., Clark, M.P., Markstrom, S.L., Viger, R.J., and Umemoto, M.: Step-wise multiple-objective calibration of a hydrologic model for a snowmelt-dominated basin, J. Am. Water Resour. A., 42(4), 877-890, doi:10.1111/j.1752-1688.2006.tb04501.x, 2006.

Hay, L.E., and McCabe, G.J.: Spatial Variability in Water-Balance Model Performance in the Conterminous United States, J. Am. Water Resour. Ass., 38(3), 847-860, DOI: 10.1111/j.1752-1688.2002.tb01001.x, 2002.

Hay, L.E., and McCabe, G.J.: Hydrologic effects of climate change in the Yukon River Basin, Climate Change, 100, 509-523, doi:10.1007/s10584-010-9805-x, 2010.

Kapangaziwiri, E., Hughes, D. A., and Wagener, T.: Constraining uncertainty in hydrological predictions for ungauged basins in southern Africa, Hydrol. Sci. J., 57, 1000–1019, 5 doi:10.1080/02626667.2012.690881, 2012.

Kiang, J.E., Stewart, D.W., Archfield, S.A., Osborne, E.B., and Eng, K.: A National Streamflow Network Gap Analysis. U.S. Geological Survey, Scientific Investigative Reports 2013-5013, 94 pp., Reston, VA, USA, 2013.

Kirchner, J.W.: Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology, J. Hydrol., 42, W03S04, doi:10.1029/2005WR004362, 2006.

Kokkonen, T.S., Jakeman, A.J., Young, P.C., and Koivusalo, H.J.: Predicting daily flows in ungauged catchments: model regionalization from catchment descriptors at the Coweeta Hydrologic Laboratory, North Carolina, Hydrol. Process., 17, 2219-2238, doi:10.1002/hyp.1329, 2003.

Krause, P., Doyle, D. P., and Bäse, F.: Comparison of different efficiency criteria for hydrological model assessment, Adv. Geosci., 5, 89–97, doi:10.5194/adgeo-5-89-2005, 2005.

Legates, D.R., and McCabe,G.J.: Evaluating the use of "goodness-of-fit" Measures in hydrologic and hydroclimatic model validation, Water Resour. Res., 35(1), 233-241, doi:10.1029/1998WR900018, 1999.

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

McCabe, G.J., Hay, L.E., Bock, A., Markstrom, S.L., and Atkinson, D.R.: Inter-annual and spatial variability of Hamon potential evapotranspiration model coefficients, J. Hydrol., 521, 389-394, doi:10.1016/j.jhydrol.2014.12.006, 2015.

McCabe, G.J., and Markstrom, S.L.: A Monthly Water-Balance Model Driven By a Graphical User Interface. U.S. Geological Survey Open-File Report 2007-1008, 12 pp., Reston, VA, USA, 2007.

McCabe, G.J., and Wolock, D.M.: Century-scale variability in global annual runoff examined using a water balance model, Int. J. Climtol., 31, 1739-1748, doi:10.1002/joc.2198, 2011a.

McCabe, G.J., and Wolock, D.M.: Independent effects of temperature and precipiation on modeled runoff in the conterminous United States, Water Resour. Res., 47, W1152, doi:10.1029/2011WR010630, 2011b.

McManamay, R.A., Orth, D.J., Dolloff, C.A., and Frimpong, E.A: Regional Frameworks applied to Hydrology: Can Landsacpes-based frameworks capture the hydrologic variability?, River Res. App., 28, 1325-1339, doi:10.1002/rra.1535, 2011.

Merz, R., and Bloschl, G.: Regionalisation of catchment model parameters, J. Hydrol., 287, 95-123, doi:10.1016/j.jhydrol.2003.09.028, 2004.

Moriasi, D.N, Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D., and Vieth, T.L.: Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations, T. ASABE, 50, 885-900, 2007.

Nash, J.E., and Sutcliffe, J.V.: River flow forecasting through conceptual models Part I: a discussion of principles, J. Hydrol., 10, 282-290, doi:10.1016/0022-1694(70)90255-6, 1970.

National Operational Hydrologic Remote Sensing Center, Snow data Assimilation System (SNODAS) Data Products at the NSIDC, 9/30/2003 through 6/13/2014, National Snow and Ice Data Center, Boulder, Colorado, USA, http://dx.doi.org/10.7265/N5TB14TC, 2004.

Newman, A.J., Clark, M.P., Sampson, K., Wood, A., Hay, L.E., Bock, A., Viger, R.J., Blodgett, D., Brekke, L., Arnold, J.R., Hopson, T., and Duan, Q.: Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance, Hydrol. Earth Syst. Sc., 19, 209-223, doi:10.5194/hess-19-209-2015, 2015.

Oudin, L., Andreassian, V., Perrin, C., Michel, C., and Le Moine, N.: Spatial proximity, physical similarity, regression and ungaged catchments: a comparison of regionalization approaches based on 913 French catchments, Water Resour. Res., 44, 1-15, doi:10.1029/2007WR006240, 2008.

Oudin, L., Kay, A., Andreassian, V., and Perrin, C.: Are seemingly physically similar catchments truly hydrologically similar?, Water Resour. Res., 46, W11558, doi:10.1029/2009WR008887, 2010.

Oyler, J.W., Dobrowski, S.Z., Ballantyne, A.P., Klene, A.E., and Running, S.W.: Artificial amplification of warming trends across the mountains of the western United States, Geophys. Res. Lett., 42, 153-161, doi:10.1002/2014GL062803, 2015.

Peel, M.C., Chiew, F.H.S., Western, A.W., and McMahon, T.A.: Extension of unimpaired monthly streamflow data and regionalization of parameter values to estimate streamflow in ungauged catchments. Report to National Land and Water Resources Audit, Center for Environmental Application and Hydrology, University of Melbourne, Parkville, 2000.

Qamar, M.U., Ganora, D., and Claps, P.: Monthly Runoff Regime Regionalization Through Dissimilarity-Based Methods, Water Resour. Manag., 29, 4735-4751, doi:10.1007/s11269-015-1087-7, 2015.

R Core Team: R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2013.

Reusser, D.: fast: Implementation of the Fourier Amplitude Sensitivity Test (FAST), R package version, http://CRAN.R-project.org/package=fast, (last access: 9 April 2014), 2012.

Reusser, D., Buytaert, W., and Zehe, E.: Temporal dynamics of model parameter sensitivity for computationally expensive models with the Fourier amplitude sensitivity test, Water Resour. Res., 47, W07551, doi:10.1029/2010WR009947, 2011.

Saltelli, A., Tarantola, S., and Campolongo, F.: Sensitivity analysis as an ingredient of modeling, Stat. Sci., 15, 377-395, 2000.

Samuel, J., Coulibaly, P., and Metcalfe, R.A.: Estimation of Continuous Streamflow in Ontario Ungauged Basins: Comparison of Regionalization Methods, J. Hydrol. Eng., 16, 447-459, http://dx.doi.org/10.1061/(ASCE)HE.1943-5584.0000338, 2011.

Sankarasubramanian, A., and Vogel, R.M.: Hydroclimatology of the continental United States, Geophys. Res. Lett., 30, 1-4, doi:10.1029/2002GL015937, 2003.

Santhi, C., Kannan, N., Arnold, J.G., and Luzio, D.: Spatial calibration and temporal validation of flow for regional scale hydrologic modeling, J. Am. Water Resour. Ass., 4, 829-846, doi:10.1111/j.1752-1688.2008.00207.x, 2008.

Sawicz, K., Wagener, T., Sivapalan, M., Troch, P.A., and Carrillo, G.: Catchment classification: empiricial analysis of hydrologic similarity based on catchment function in the eastern USA, Hydrol. Earth Syst. Sc., 15, 2895-2911, 2011.

Sefton, C.E.M., and Howarth, S.M.: Relationships between dynamic response characteristics and physical descriptors of catchments in England and Wales, J. Hydrol., 211, 11-16, doi:10.1016/S0022-1694(98)00163-2, 1998.

Seibert, J.: Regionalization of parameters for a conceptual rainfall runoff model, Agr. Forest Meteorol., 98-99, 279-293, doi:10.1016/S0168-1923(99)00105-7, 1999.

Smakhtin, V.Y.: Low flow hydrology: a review, J. Hydrol., 240, 147-186, doi:10.1016/S0022-1694(00)00340-1, 2001.

Tang, Y., Reed, P., Wagener, T., and van Werkhoven, T.: Comparing sensitivity analysis methods to advance lumped watershed model identification and evaluation, Hydrol. Earth Syst. Sc., 11, 793-817, 2007.

Tekleab, S., Uhlenbrook, S., Mohamed, Y., Savenije, H.H.G., Temesgen, M., and Wenninger, J.: Water balance modeling of Upper Blue Nile catchments using a top-down approach, Hydrol. Earth Syst. Sci., 15, 2179-2193, doi:10.5194/hess-15-2179-2011, 2011.

Troch, P.A., Paniconi, C., and McLaughlin, D.: Catchment-scale hydrological modeling and data assimilation, Adv. Water Resour., 26, 131-135, doi:10.1016/S0309-1708(02)00087-8, 2003.

USEPA (United States Environmental Protection Agency) and USGS (United States Geological Survey): NHDPlus User Guide, available at *ftp://ftp.horizon-systems.com/NHDPlus/documentation/NHDPLUS_UserGuide.pdf* (last access 12 Nov 2014), 2010.

US Geological Survey: A National Water Information System, available at: http://waterdata. usgs.gov/nwis/ (last access 27 March 2014), 2014.

Van Griensven, A., Meixner, T., Grunwald, S., Bishop, T., Diluzio, and M., Srinivasan, R.: A global sensitivity analysis tool for the parameters of multi-variable catchment models, J. Hydrol., 324, 10-23, doi:10.1016/j.jhydrol.2005.09.008, 2006.

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

Viger, R., Bock, A.: GIS Features of the Geospatial Fabric for National Hydrologic Modeling, U.S. Geological Survey, Denver, CO, USA, doi:10.5066/F7542KMD, 2014.

Vogel, R.M.: Regional calibration of watershed models, Watershed Models, Singh, V.P., and Frevert, D.F. (Eds.), CRC Press, Boca Raton, FL, USA, 2006.

Vrught, J.A., ter Braak, C.J.F., Clark, M.P., Hyman, J.M., Robinson, B.A.: Treatment of input uncertainty in hydrologic modeling: Doing hydrology backwards with Markov Chain Monte Carlo simulation, Water Resour. Res., 44, W00B09, doi:10.1029/2007WR006720, 2008.

Wolock, D.M.: STATSGO soil characteristics for the conterminous United States. U.S. Geological Survey Open-File Report 1997-656, Reston, VA, USA, available at: *http://water.usgs.gov/GIS/metadata/usgswrd/XML/muid.xml*, (last access 3 March 2012), 1997.

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

Zhang, X., Srinivasan, R., and Van Liew, M.: Multi-Site Calibration of the SWAT Model for Hydrologic Modeling, T. ASABE, 51, 2039-2049, 2008.

Parameter	Definition	Range	Default
1. Drofac	Controls fraction of precipitation that becomes runoff	0, 0.10	0.05
2. Rfactor	Controls fraction of surplus that becomes runoff	0.10, 1.0	0.5
3. Tsnow	Threshold above which all precipitation is rain (^{o}C)	-10.0, -2.0	-4.0
4. Train	Threshold below which all precipitation is snow (^{o}C)	0.0, 10.0	7.0
5. Meltcoef	Proportion of snowpack that becomes runoff	0.0, 1.0	0.47
6. Ppt_adj	Seasonal adjustment factor for precipitation (%)	0.5, 2.0	1
7. Tav_adj	Seasonal adjustment for temperature (^{o}C)	-3.0, 3.0	0

Table 1. Monthly Water Balance Model parameters and ranges.

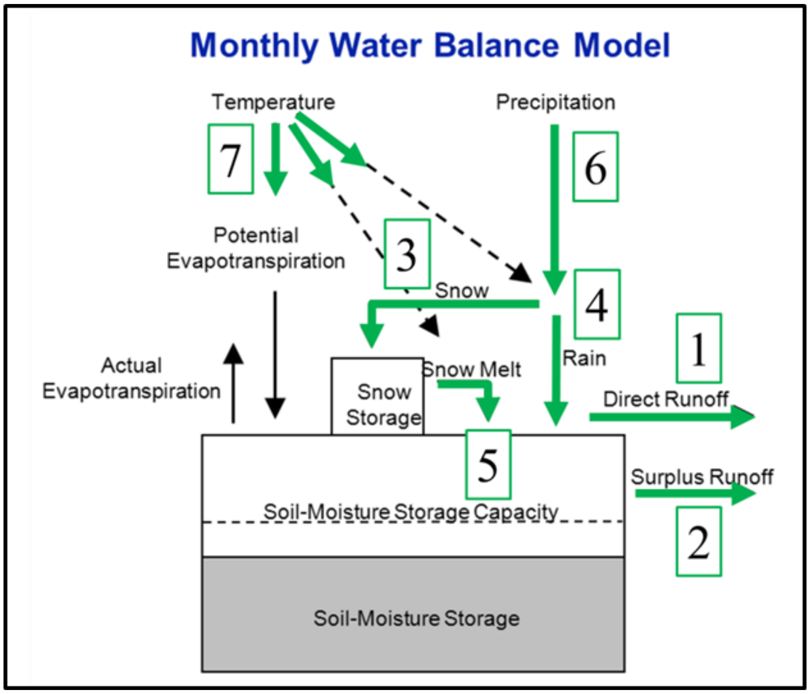

Figure 1. Conceptual diagram of the Monthly Water Balance Model (McCabe and Markstrom 2007). Processes influenced by model parameters used in Fourier Amplitude Sensitivity Test (FAST) are identified by green arrow and numbered (Table 1).

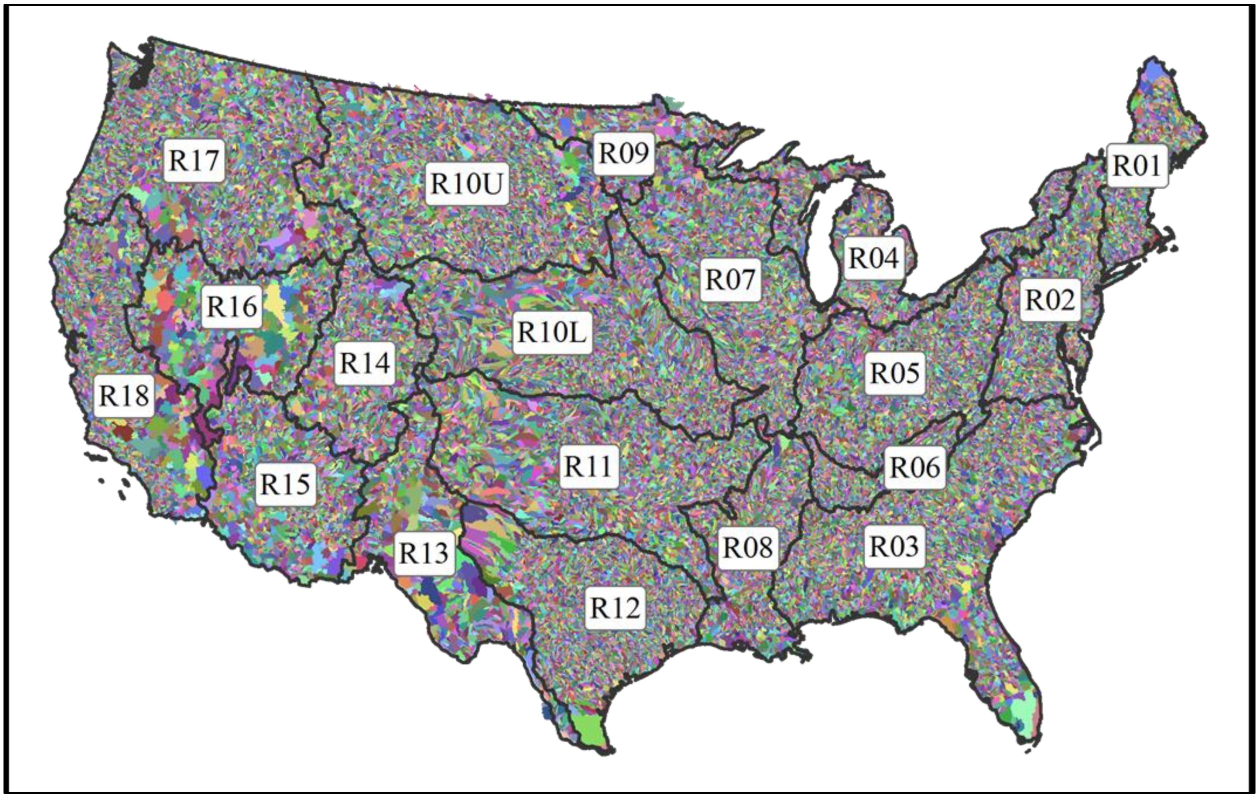

Figure 2. Hydrologic Response Units of the Geospatial Fabric, differentiated by color, overlain by NHDPlus region boundaries (R01-R18).

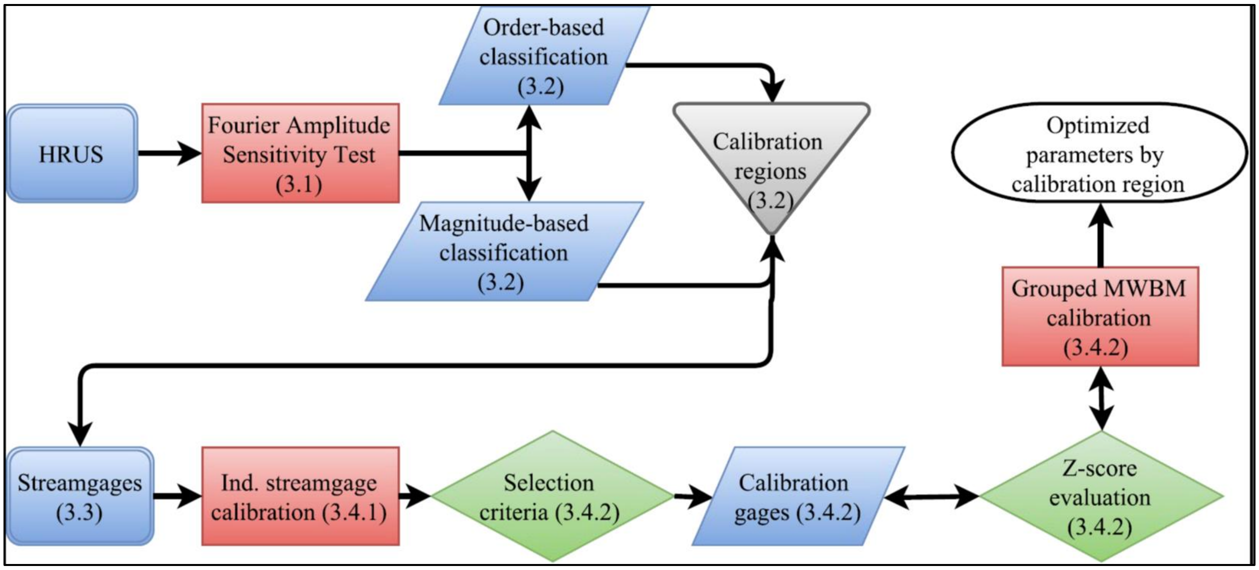

Figure 3. Schematic flowchart of the parameter sensitivity analysis and regionalization method described in this paper (Section 3).

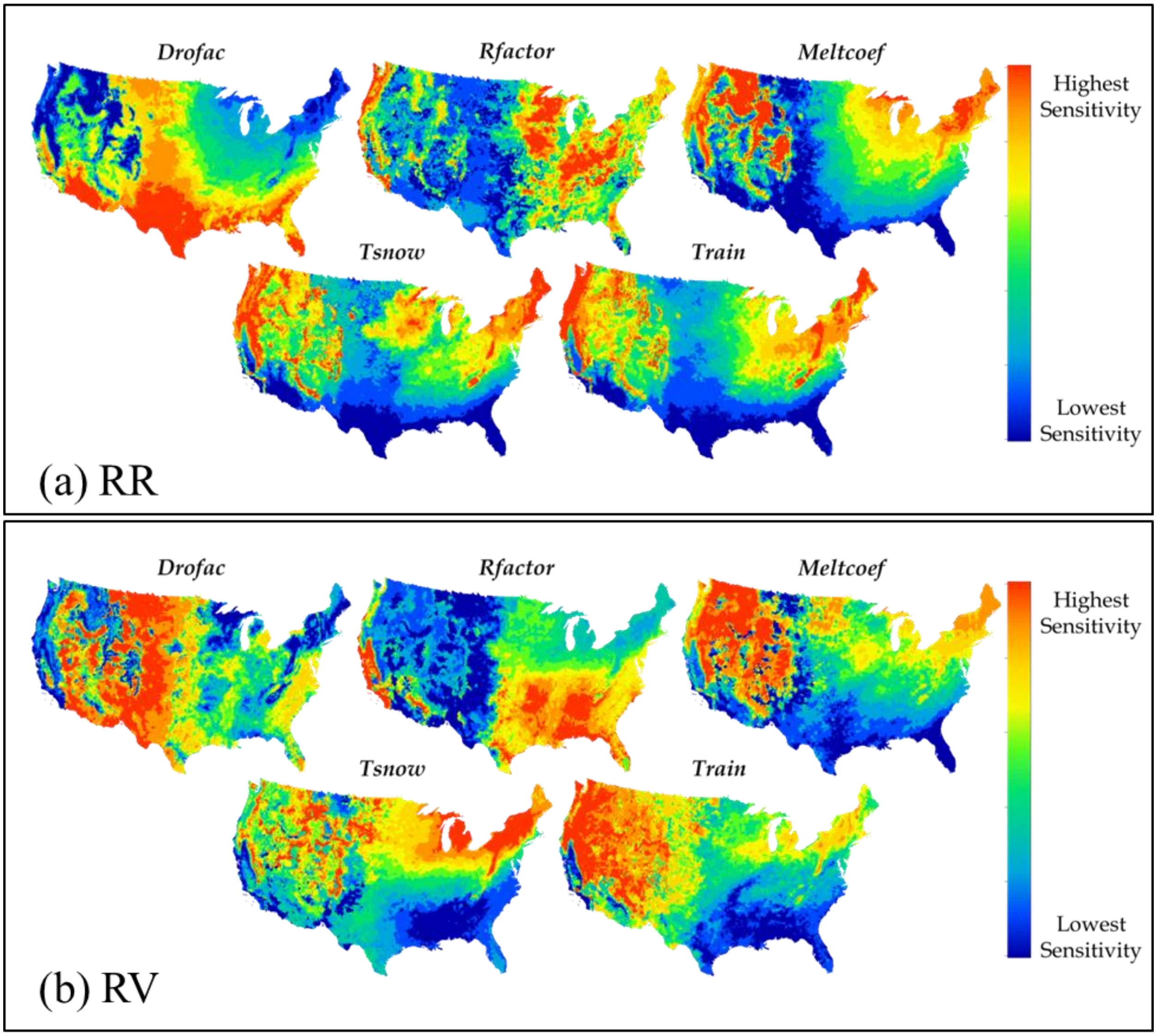

Figure 4. Relative sensitivity of the (a) Rainfall Ratio (RR) and (b) Runoff Variability (RV) indices to Monthly Water Balance Model parameters.

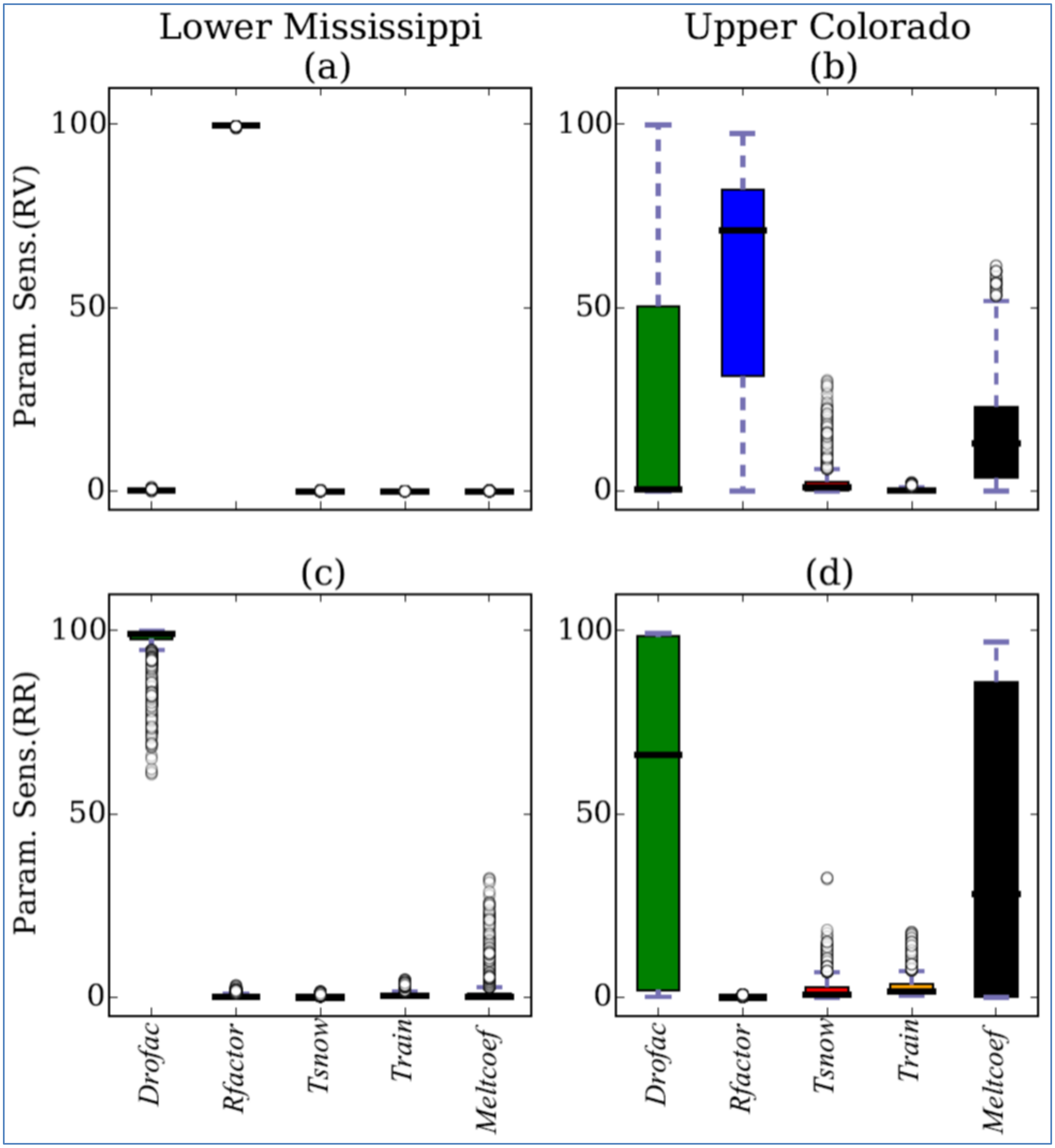

Figure 5. Parameter sensitivities of Runoff Variability (RV; a and b) and Runoff Ratio (RR; c and d) indices for Monthly Water Balance Model parameters in the Lower Mississippi (R08) and Upper Colorado (R14).

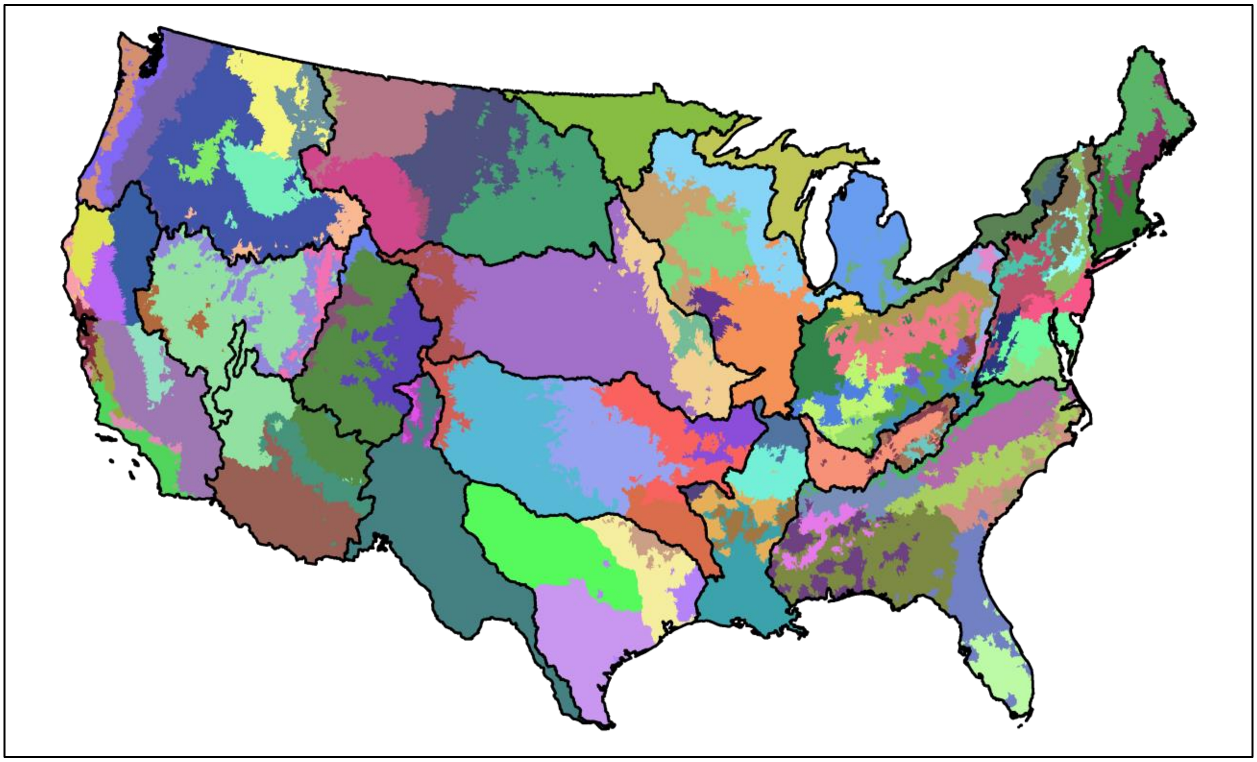

Figure 6. Final 110 Monthly Water Balance Model calibration regions differentiated by colors. A subset of streamgages within each calibration region were calibrated in a group-wise fashion to produce a single optimized parameter set for the entire region (Fig. 3).

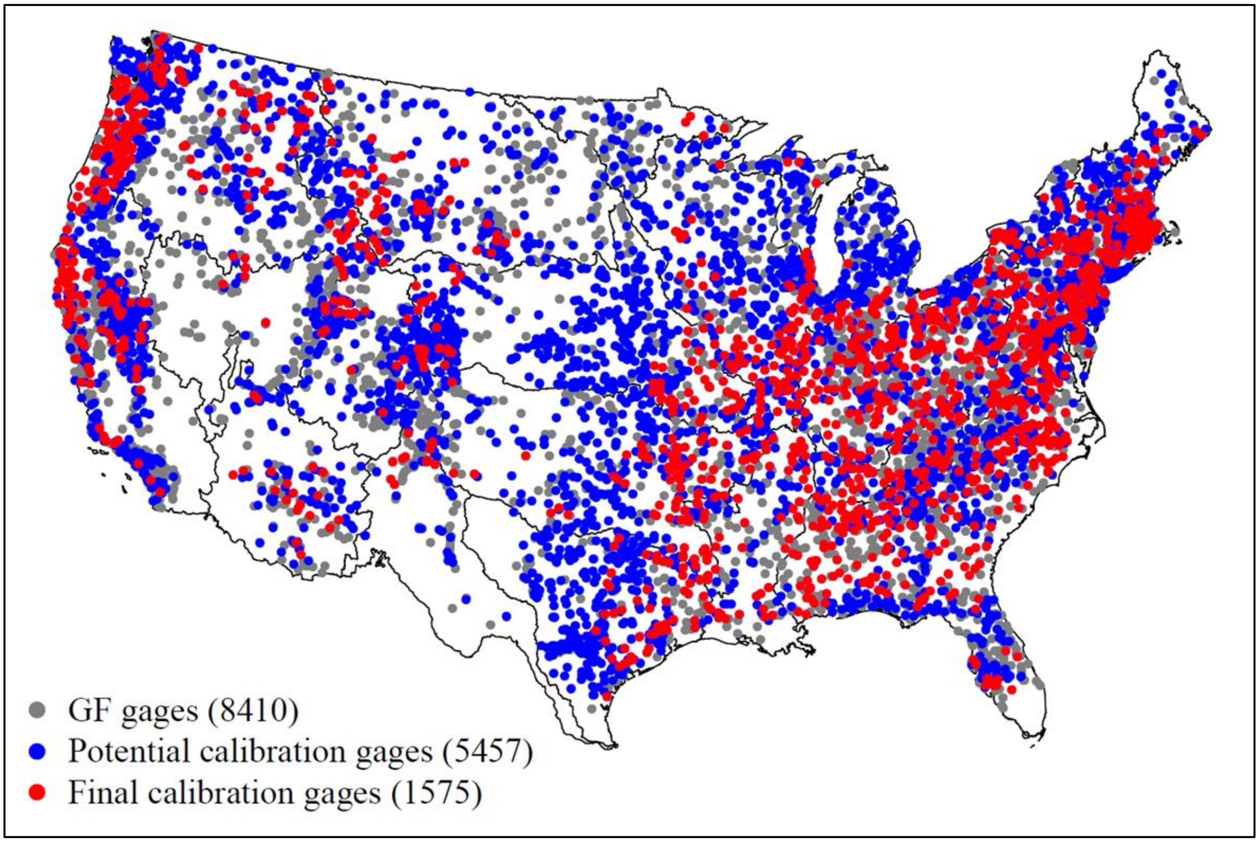

Figure 7. Streamgages tested in the study. GF notes geospatial fabric for national hydrologic modeling (Viger and Bock, 2014).

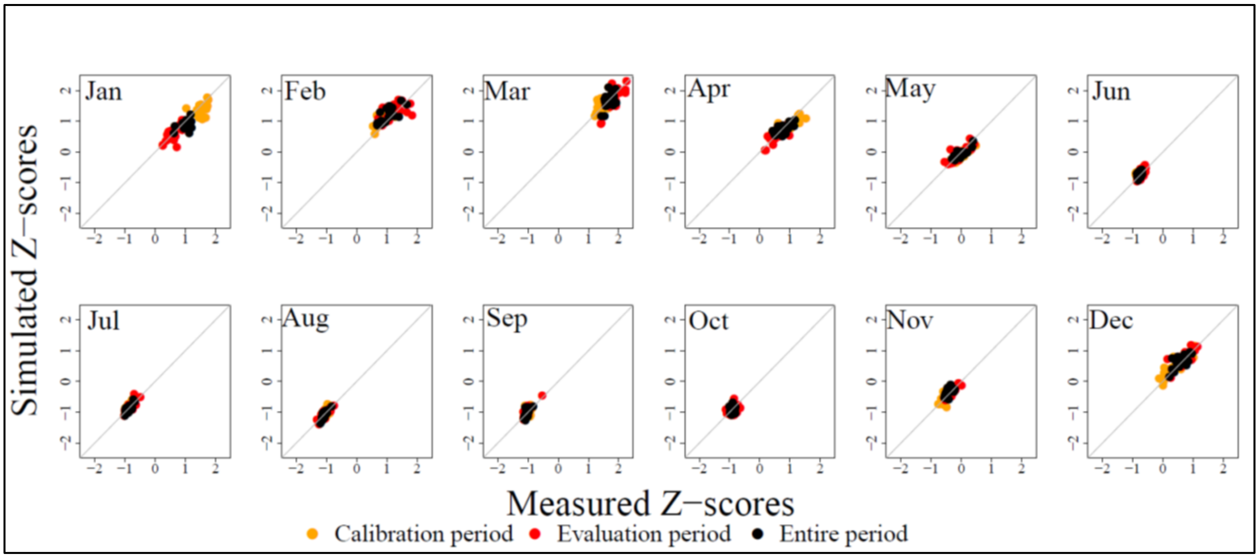

Figure 8. Measured versus simulated mean monthly Z-scores for the Tennessee River calibration region (see Fig. 10b for location). Orange is calibration, red is evaluation, and black is all years.

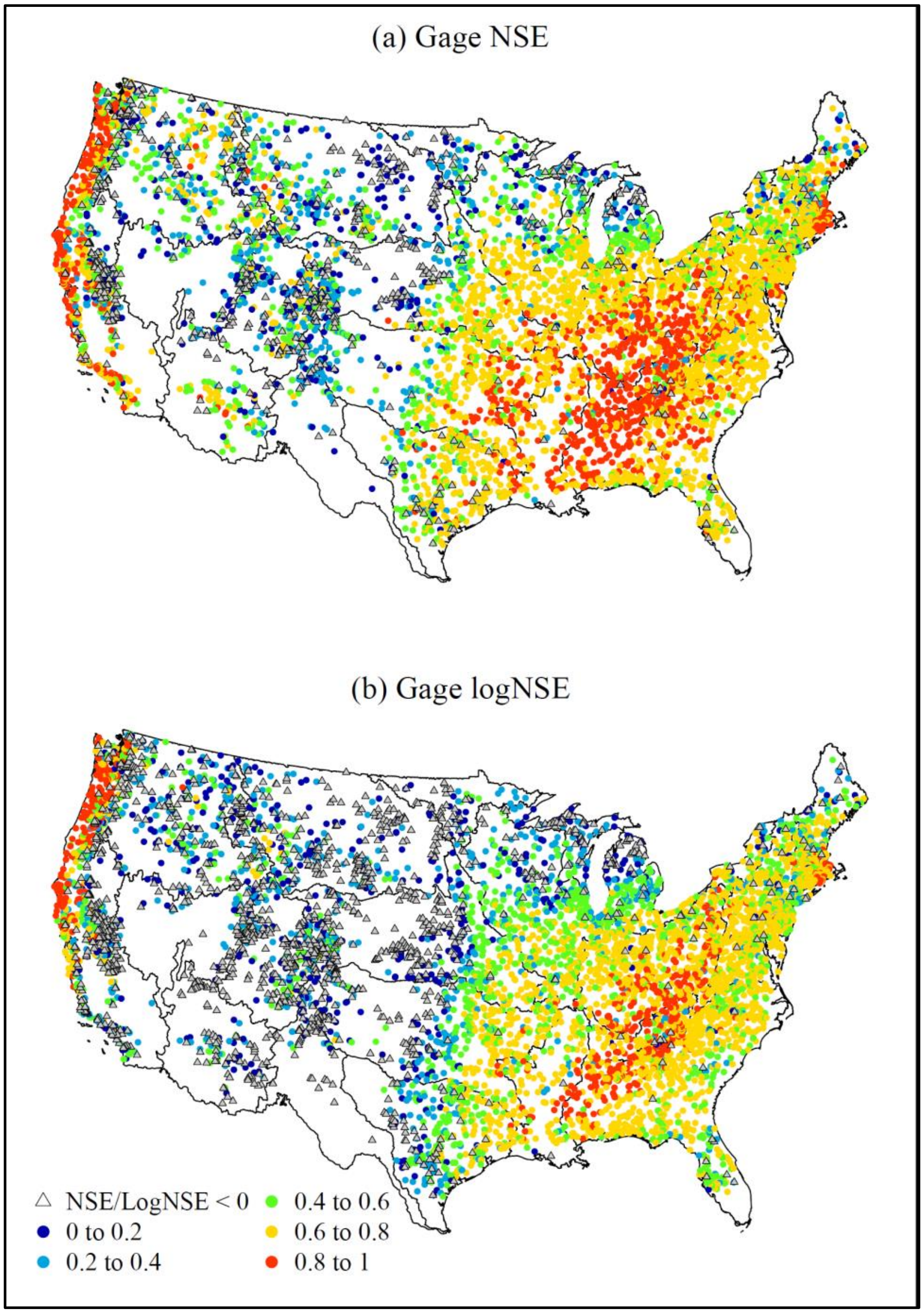

Figure 9. Individual streamgage calibration results: (a) Nash-Sutcliffe Efficiency (NSE) coefficient and (b) log of the NSE (logNSE).

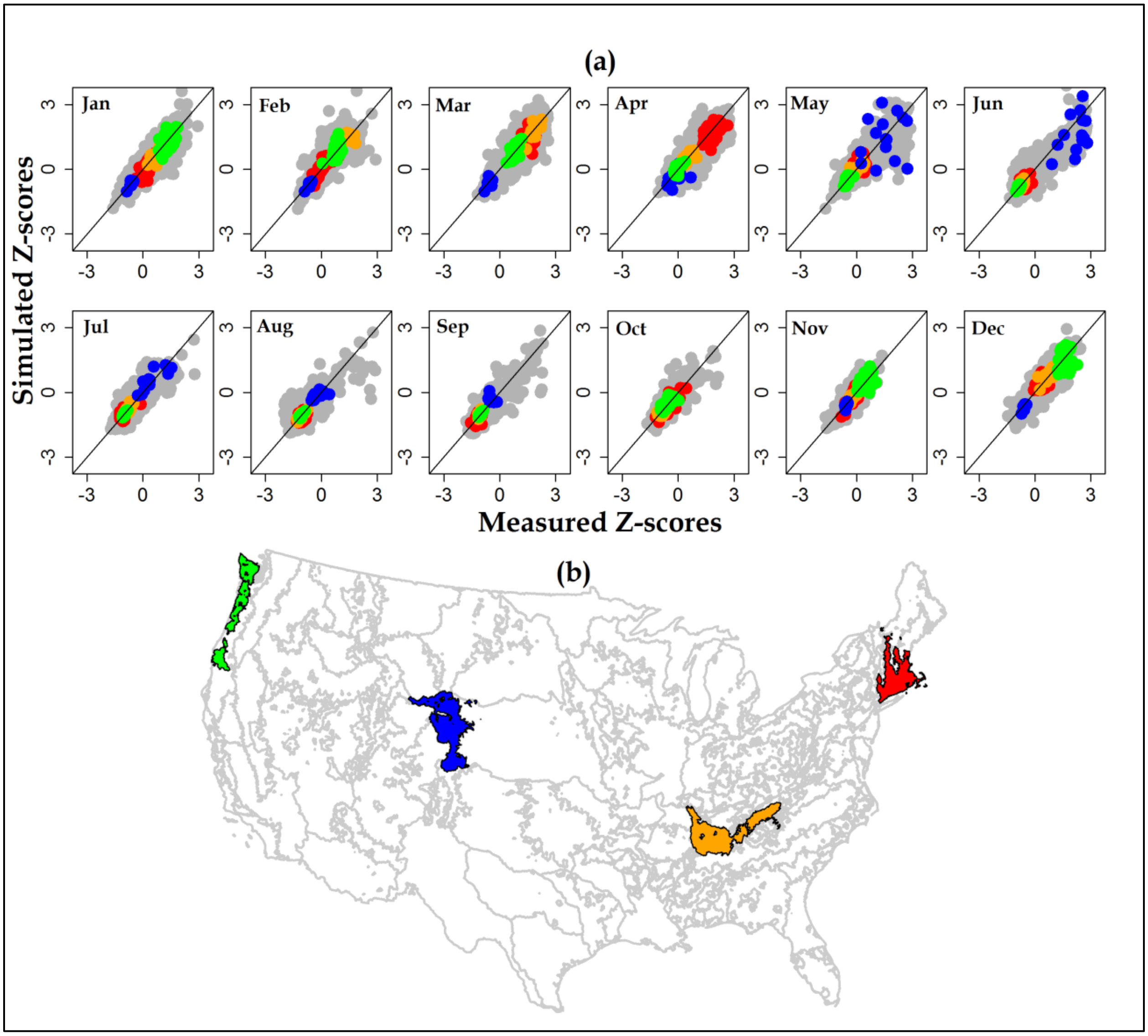

Figure 10. (a) Measured versus simulated mean monthly Z-scores for runoff at all streamgages and (b) location of highlighted streamgages for four calibration regions: New England (67 streamgages, red); Tennessee River (21 streamgages, orange); Platte Headwaters (15 streamgages, blue); and Pacific Northwest (33 streamgages, green).

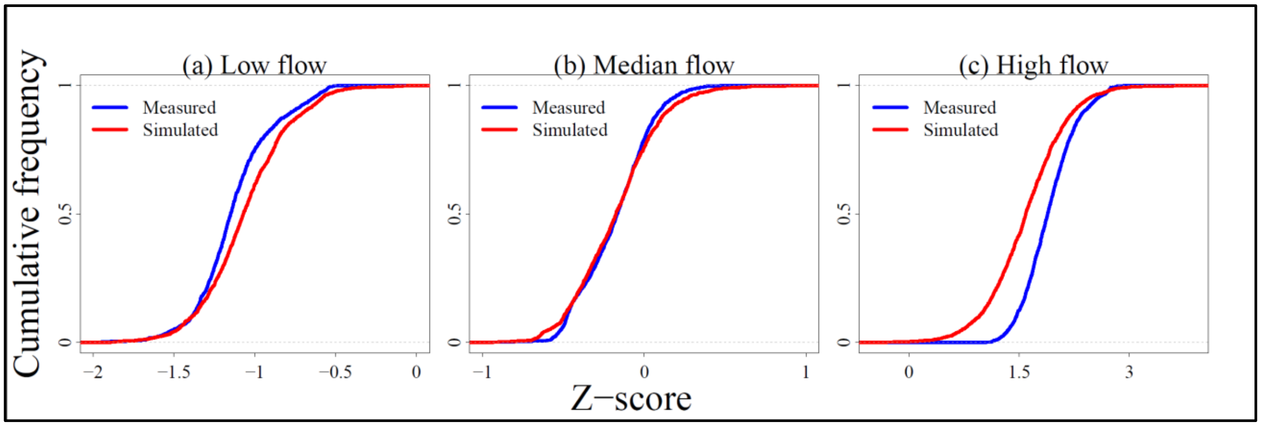

Figure 11. Z-score cumulative frequency for (a) highest-, (b) median-, and (c) lowest-flow months.

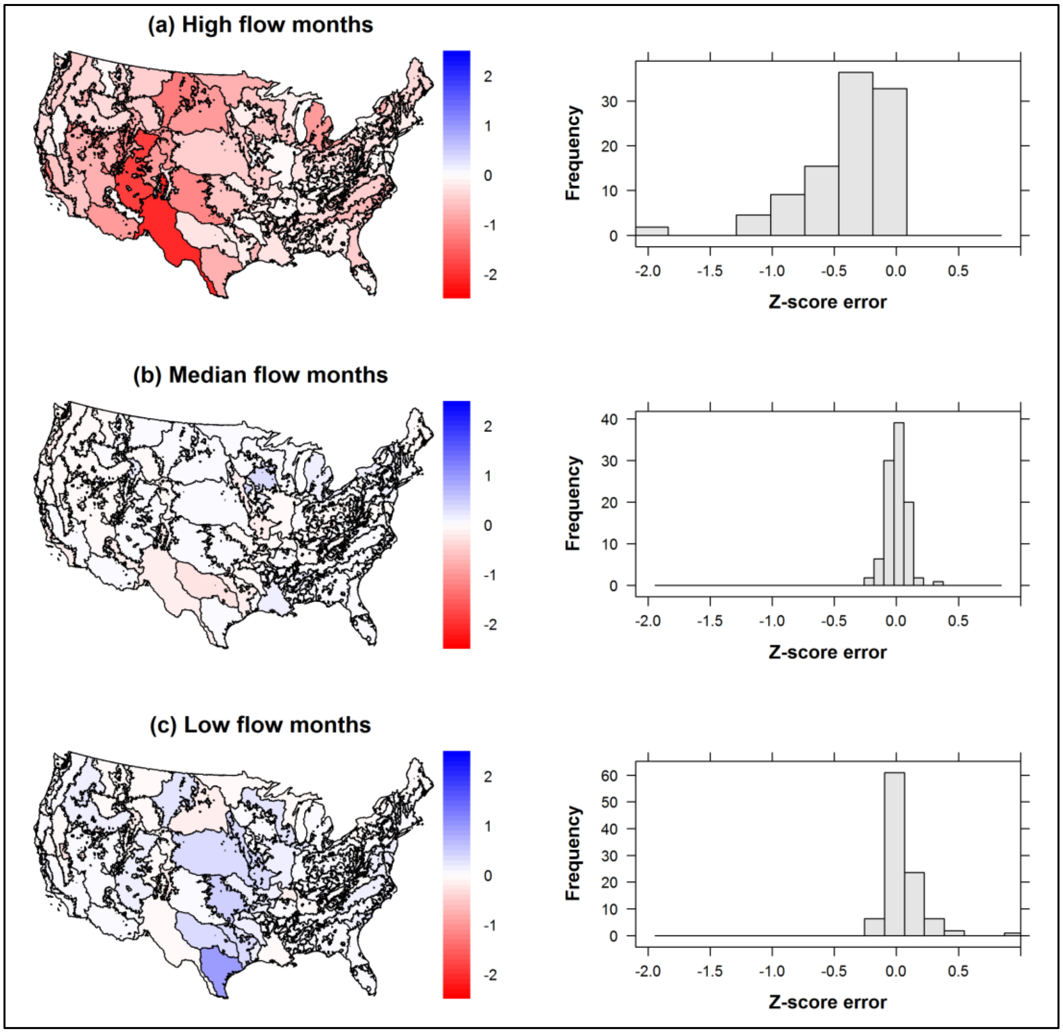

Figure 12. Z-score error (simulated - measured) for (a) highest-, (b) median-, and (c) lowest-flow months.

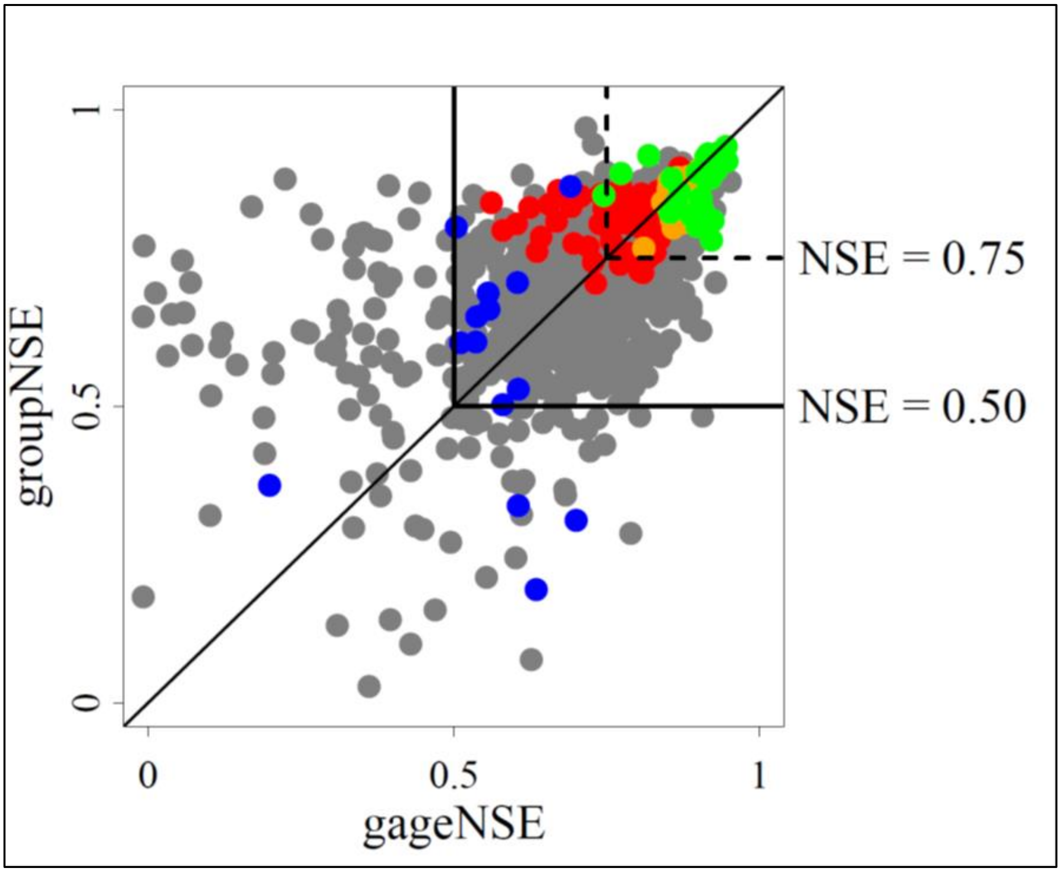

Figure 13. Nash Sutcliffe Efficiency from individual (gageNSE) and grouped (groupNSE) calibration. Calibration regions in New England (67 streamgages, red); Tennessee River (21 streamgages, orange); Platte Headwaters (15 streamgages, blue); and Pacific Northwest (33 streamgages, green) are highlighted (see Fig. 10b for location).

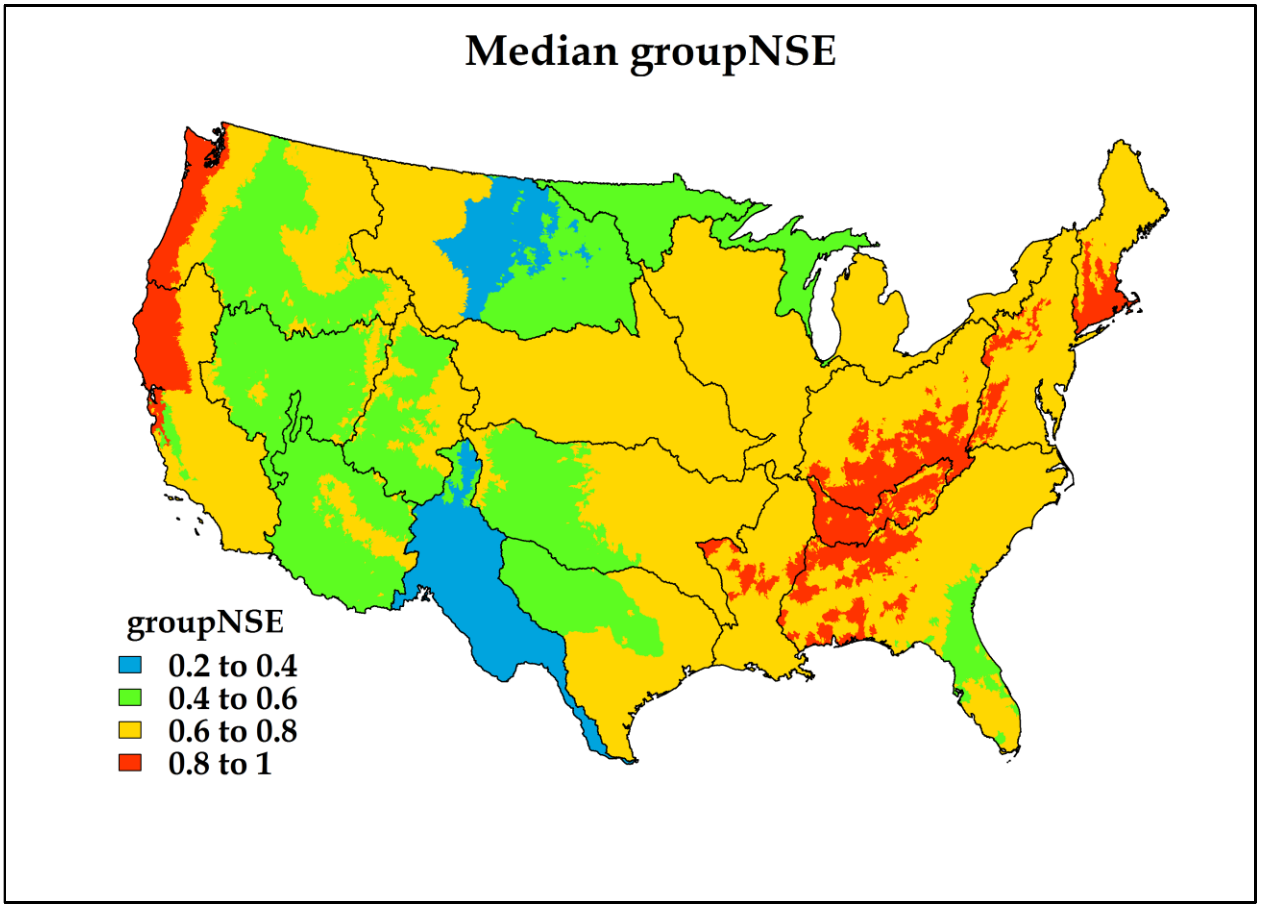

Figure 14. Median Nash Sutcliffe Efficiency (NSE) by calibration region of streamgages used for calibration.