Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-121, 2017
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 6 March 2017

Hydrology and Earth System Sciences
Open Access
EGU
Discussions

Multiple domain evaluation of watershed hydrology models

Karthik Kumarasamy[1], Patrick Belmont[1]

[1]Department of Watershed Sciences, Utah State University, Logan, 84322-5210, USA

Correspondence to: Karthik Kumarasamy (karthik.k@aggiemail.usu.edu)

Abstract. Watershed scale models simulating hydrology and water quality have advanced rapidly in sophistication, process representation, flexibility in model structure, and input data. Given the importance of these models to support decision-making for a wide range of environmental issues, the hydrology community is compelled to improve the metrics used to evaluate model performance. More targeted and comprehensive metrics will facilitate better and more efficient calibration and will help demonstrate that the model is useful for the intended purpose. Here we introduce a suite of new tools for model evaluation, packaged as an open-source Hydrologic Model Evaluation (HydroME) Toolbox. Specifically, we demonstrate the use of box plots to illustrate the full distribution of common model performance metrics, such as R^2, use of Euclidian distance, empirical Quantile-Quantile (Q-Q) plots and flow duration curves as simple metrics to identify and localize errors in model simulations. Further, we demonstrate the use of magnitude squared coherence to compare the frequency content between observed and modeled streamflow and wavelet coherence to localize frequency mismatches in time. We provide a rationale for a hierarchical selection of parameters to adjust during calibration and recommend that modelers progress from parameters with the most uncertainty to the least uncertainty, namely starting with pure calibration parameters, followed by derived parameters, and finally measured parameters. We apply these techniques in the calibration and evaluation of models of two watersheds, the Le Sueur River Basin (2880 km^2) and Root River Basin (4300 km^2) in southern Minnesota, USA.

1 Introduction

1.1 Hydrologic models and the calibration challenge

Watershed scale models simulating hydrology and water quality have evolved considerably over the past few decades. Such models are essential to inform policy and management decisions at scales ranging from individual farm fields to the entire Mississippi River Basin (Santhi et al., 2014). Leading models, such as the Soil and Water Assessment Tool (SWAT), Water Erosion Prediction Project (WEPP) and Gridded Surface Subsurface Hydrologic Analysis (GSSHA) have advanced rapidly in sophistication, process representation and flexibility regarding model components and input data. The caveat that accompanies this sophistication and flexibility is that each process is often represented by multiple adjustable parameters, which greatly increases the challenge of calibration and problem of equifinality. In the spirit of continued advancement, this paper examines current techniques and metrics used to evaluate hydrologic models for the purpose of calibration and validation and proposes several additional techniques to improve quantitative evaluation of models.

Calibration is the process of estimating model parameter values to enable a hydrologic model to match observations (Singh and Frevert, 2002) such as streamflow. Some form of calibration is necessary for most distributed hydrologic model applications to account for limitations in model structure, data availability, and initial and boundary conditions ((Beven, 2011) and references therein). Over the past few decades, numerous publications have highlighted the challenges associated with calibration such as; physical distortion by tuning incorrect parameters, inability of performance measures used in calibration to capture all aspects of the hydrologic times series, and limitations in the search schemes, such as Monte Carlo Markov Chain, uniform Monte Carlo and Latin Hypercube (Sorooshian and Gupta, 1983;Beven, 2006;Madsen, 2003;Beven and Freer, 2001). Several techniques have been proposed to partially resolve some of these challenges (Singh and Frevert, 2002;Beven, 2011). While the theoretical underpinnings of calibration have advanced considerably (Gupta et al., 1998), several key challenges persist. Challenges stem from the sheer number and complexity of processes involved, many of which are non-linear, exhibit interdependencies, and are subject to large variability in both time and space. Automatic calibration procedures have become common for watershed hydrology models as manual calibration can be somewhat inefficient and time consuming.

Gathering more and higher resolution data will not fully resolve the calibration issues as many parameters are conceptual representations of abstract physical processes and therefore cannot be measured (Singh and Frevert, 2002). In this paper we focus on (1) techniques to identify offsets in the magnitude and frequency of modeled compared to measured streamflows (2) hierarchical prioritization of parameters to adjust during calibration and (3) moving beyond lumped time domain metrics for model evaluation to consider hydrograph shape and signal frequency. We introduce several new tools to improve model evaluation, packaged as the Hydrologic Model Evaluation (HydroME) Toolbox, which is freely available for download from https://qcnr.usu.edu/labs/belmont_lab/resources.

We utilize SWAT throughout this paper because it has emerged as a leading model for informing policy and management. SWAT benefits from an enormous user and developer base that has struck a synergistic balance of encouraging grassroots innovation and adaptation (e.g., (Chen and Wu, 2012)), while maintaining model stability and version control. In many ways, it is a Community Hydrologic Model called for by Weiler and Beven (2015). In 2015 alone, 437 journal articles were published based on the SWAT model (CARD and ISU of Science and Technology, 2016). However, we note that the tools and approach we take are more broadly applicable to any watershed hydrology model.

1.2 Model evaluation and reporting measures

We propose that the discipline of hydrologic modeling is sufficiently mature to adopt a new suite of model performance metrics that more specifically and meaningfully convey the suitability of the model to answer the questions of interest. The increasingly interdisciplinary nature of hydrology and wide-ranging use of hydrologic models to make predictions about water quality, sediment transport, ecological processes and ecosystem health increase the urgency for more targeted and robust measures of model performance. In effect, we recognize the wise sentiment that all models are wrong in an absolute sense and that it is up to the modeler to provide compelling evidence that the model is useful for the intended purpose, which increasingly requires much more than simple goodness-of-fit performance metrics lumped at monthly or annual timescales.

Lumped metrics such as Nash Sutcliff Efficiency (NSE) and coefficient of determination (R^2) have been established as key model performance benchmarks. These metrics provide an averaged measure of error and are intentionally biased towards large magnitude flows (Criss and Winston, 2008). NSE is slightly better than R^2 for many model applications as it is sensitive to the observed and model simulated means and variances (Krause et al., 2005). It is important to note, however, that these metrics only address magnitude errors and are insensitive to critical flow thresholds that may be important to answer the questions of interest. Many other streamflow characteristics may be important depending on whether your model is being used to simulate water, sediment or nutrient fluxes or aquatic habitat quality (Lytle and Poff, 2004;Sanborn and Bledsoe, 2006;Wenger et al., 2010). Increasingly, watershed hydrology models are used for predictions, or as inputs to other models with processes operating at daily and sub-daily time steps. In such cases, models evaluated at monthly scales are of limited value. Further, streamflow event structures are not characterized by these metrics. For example, NSE and R^2 are ambiguous as to shape of streamflow hydrograph from an individual storm. The goal of this paper is to provide insight and new tools that facilitate meaningful model evaluation throughout the process of calibration, validation and communication of results.

2 Study Area

We illustrate our approach using two carefully selected case studies. The contrasting environments represented by our study watersheds challenge the model structure in different ways and present distinct calibration challenges.

2.1 Le Sueur River Basin (LSRB)

The 2880 km^2 Le Sueur River Basin in south-central Minnesota is listed as impaired for excessive sediment and nutrients and is implicated as a primary contributor to water quality problems in the Minnesota River and Lake Pepin on the Mississippi River (Wilcock and Belmont, 2009;Belmont et al., 2011;WRC and MPCA, 2009). Flow and water quality have been monitored at 8 locations (23-2880 km^2) throughout the watershed and the modern water quality problems have been shown to be a combination of natural and human factors (Belmont, 2011;Gran et al., 2011), with river channel erosion playing a dominant role in sediment loading (Belmont et al., 2014). The landscape is quite flat except for the lower 40 km of the mainstem and two major tributaries (the Big Cobb and Maple rivers) (Belmont et al., 2011). The lower 40 km of all three rivers are rapidly incising due to a 70 m base level fall caused by a catastrophic glacial outwash event that carved out the Minnesota River Valley 13,400 years ago (Thorleifson, 1996;Fisher, 2003;Belmont, 2011;Gran et al., 2013). Recent research has highlighted the fact that flows have increased considerably, due to increased precipitation and enhanced agricultural drainage practices (Schottler et al., 2014). It is essential that hydrologic models developed to inform watershed conservation and restoration actions are well calibrated and do not physically distort the system. Achieving this goal is complicated by the complex sub-surface hydrology of the Le Sueur, which contains a sequence of highly heterogeneous tills and glaciofluvial sediments (Jennings, 2010). In addition, humans have profoundly altered the sub-surface hydrology with an extensive network of drain tiles (corrugated plastic tubing installed 30-100 cm below the soil surface).

ArcSWAT 2012.10.1.15 for ArcGIS 10.1 was used to extract SWAT friendly text files using 10 m DEM for topography (USGS, 2013a, b), Cropland Data Layer for land use (Weiguo et al., 2014) and SSURGO data for soils (Soil Survey Staff, 2015). Temperature (maximum and minimum) and precipitation data were obtained from (PRISM Climate Group, 2004) at 4 km resolution and averaged daily within each of the 175 sub-basins. Solar radiation and relative humidity data were obtained from global weather data for SWAT (Saha et al., 2014). Hydrologic Response Units (HRUs) were defined using the multiple HRUs option with 5% for land use, 15% for soil and 10% for slope. Corn, soybean and wetlands were exempted from the land use threshold definition resulting in 1823 HRUs for the basin. Management practices implemented in the model are shown in Fig. S1 in Supplemental Information (SI).

A multipoint and multi-parameter calibration was employed and the model was calibrated and validated against daily streamflow at 8 gages within the basin. Parameters selected for calibration and their calibrated values are listed in Table S2 in SI. Streamflow measurements at upstream gages were calibrated first, followed by downstream gages using three goodness of fit metrics: (1) NSE, (2) R^2, (3) PBIAS in addition to other metrics discussed below. The relevance of these metrics to this study is further described in the SI.

2.2 Root River basin (RRB)

The 4300 km^2 Root River Basin in southeastern Minnesota is also listed as impaired for excess sediment and nutrients under the USEPA Clean Water Act (MPCA, 2012). The upper third of the watershed is flat and underlain by fine-grained glacial till, similar to the unincised portions of the Le Sueur watershed. The lower two-thirds of the basin are within the 'Driftless Area' region, which has not been glaciated for the past 500,000 years (Knox, 1987;Troelstrup and Perry, 1989). The majority of this portion of the watershed is dominated by Paleozoic limestone and dolostone bedrock with hundreds of caves and sinkholes connecting surface flowpaths to the poorly mapped karst groundwater network. Topography of this zone is characterized by relatively steep, forested hillslopes with row crop agriculture and pasture on lower sloped terrain. The lower reaches of the mainstem and major tributaries are distinctly characterized by low gradient, unconfined alluvial river valleys that have been aggrading throughout the Holocene as the mainstem Mississippi River aggraded from late-Pleistocene sedimentation. Pending watershed conservation and restoration decisions require reliable information, part of which must be extracted from a well calibrated hydrologic model (Stout et al., 2014;Belmont et al., 2016a). Further, the geologic, geomorphic and hydrologic setting of the Root River provides a useful contrast to the Le Sueur watershed for our discussion of model calibration and evaluation.

The RRB SWAT model uses the SWAT2012.exe Revision 622 packaged with ArcSWAT 2012.10_1.15. Subbasin delineation is accomplished using 10m topography data, by choosing a threshold-based stream definition and by specifying locations where a model output is required to facilitate comparison with measured streamflow. The threshold based streamflow definition option allows us to control the size of the subbasins, which is crucial for capturing variability in precipitation and also to simulate flows at any desired location where flow outputs are needed. The RRB model was built using the multiple HRUs definition option by specifying a threshold of 20% for land use, 0% for soils, and 0% for slope, resulting in 17,174

HRUs. Topography, soils, and land use data were obtained from same sources as Le Sueur River basin. Management practices implemented are shown in Fig. S2 in SI. All the model parameters that were calibrated are listed in Table S3 in SI.

Karst presents a model structure limitation as SWAT does not explicitly model preferential flow pathways. However, SWAT is commonly applied to such watersheds (e.g., (Baffaut and Benson, 2009)) with lumped treatment of such flow pathways. RRB model was initialized with land use data from 2006, which is the first year when the USDA Cropland Data Layer (CDL) was available for Minnesota. Details on how this information was compiled is described in Belmont et al. (2016a). Hydrologic effects of the karst system (i.e., stream flow loss or gain) were accomplished by altering tributary and main channel transmission losses. Rapid and slow responses of groundwater contribution to streamflow that result from preferential pathways were accomplished by altering groundwater delay times and rate and quantity of groundwater that is fed to streamflow. Similar to the LSRB SWAT model, a multipoint and multi-parameter calibration was employed and the model was calibrated and validated against daily streamflow at 5 gages within the basin. Parameters selected for calibration and their calibrated values are listed in Table S3 in SI.

3 Parameter choice, estimation, and classification

3.1 Parameter choice problem

Choosing which parameters to adjust in a complex watershed hydrology model such as SWAT can be a daunting task. Any given parameter can be sensitive within a particular range and during specific seasons or time periods. Further, the sensitivity of any given parameter may vary depending on one or multiple other parameter values. How should one choose which parameters to adjust in a complex model such as SWAT, with over 100 potentially adjustable parameters to choose from? More importantly, when one achieves what is deemed to be a good or acceptable calibration, how can they be confident that they have achieved high performance metrics for the right reasons and have not mathematically distorted the physical system in ways that may bias their results or interpretations? Multiple model parameterizations can result in similar modeled streamflow outcomes, which may be completely different representations of the physical systems. Motivated by these shortcomings implicit in the blind (nondiscriminatory) tuning of parameters we propose some general guidelines for selection of parameters.

Prior to adjusting parameters for calibration it is prudent to define the range of reasonable values from literature that employ similar models in similar watersheds (Singh and Frevert, 2002). While we encourage use of this approach, we also acknowledge the inherent assumption that the published parameter values were obtained in a robust manner and note the inevitable limitation that no two watersheds are identical. Further, it is important to note that different model structures can substantially alter simulation results (Butts et al., 2004). For example, SWAT provides three evapotranspiration (ET) and two infiltration excess runoff structures (Neitsch et al., 2011), each of which have distinct implications on parameters related to runoff generation. While choosing the best structure for your input data and application is an important step in the calibration process, exploring the effects of different model structures goes beyond the scope of this paper.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-121, 2017
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 6 March 2017

3.2 Fully automated and intervention type calibration procedures

Automated calibration and other hybrid procedures have become increasingly common over the past decade (Madsen et al., 2002;Tolson and Shoemaker, 2007;Efstratiadis and Koutsoyiannis, 2010). The objective nature of automated parameter adjustment makes these approaches very appealing. And indeed, these approaches offer many advantages and will become increasingly effective as techniques and algorithms improve. However, these procedures typically rely, often exclusively, on a single lumped closeness measure, such as Root Mean Square Error (RMSE) or the R^2. As described in Singh and Frevert (2002) and discussed above, a single criterion is inherently predisposed to bias towards certain components of the hydrologic time series. Automated procedures will be improved by the use of multiple evaluation criteria as discussed here and by Gupta et al. (1998). However, automated procedures adjust parameter values without regard for process or physical meaning of any given combination. Parameters are adjusted based on search schemes that generally seek to minimize the number of runs or explore the parameter space to maximize a given, typically time-lumped, performance metric (Beven and Freer, 2001). Thus, they may adjust some parameters erroneously. They can also be limited by the number of iterations or through convergence in terms of a stopping criterion (Madsen, 2003). Although automation can help the calibration process become more objective, efficient and practical, these procedures are not a panacea substitute for expert hydrologic intuition and understanding.

Problems of equifinality, determining when a model is sufficiently well calibrated, and interdependencies among parameters present formidable challenges in calibration. For example, Fig 2 (a) shows the contour plot of NSE values computed from 1000 model runs, varying two flow routing parameters. Specifically, we varied Manning's "n" value for the tributary channels (CH_N1) between 0.025-0.035 and varied Manning's "n" value for the main channel (CH_N2) between 0.025-0.033, which are reasonable for “short grass” and “clean, straight, full stage, no rifts or deep pools,” respectively (Chow, 1959). Desirable NSE values, which we define as NSE values > 0.6 occur in the parameter space > 0.026 for CH_N2 and > 0.03 for CH_N1. However, if we hold tributary channel roughness (CH_N1) constant at the “normal” value of 0.03 and instead vary the Curve Number (CN2) and Manning's "n" value for the main channel (CH_N2), we can see that the entire space of CH_N2 yields desirable NSE values (i.e., NSE > 0.6 across the range of CH_N2 values for all negative adjustments in CN2 in Fig. 2 (b)). Further, when we alter 18 parameters which are listed in Table S1 in SI, we now see that the entire parameter space for the both CH_N1 and CH_N2 contains desirable solutions (see dark blue portions of Fig. 2 (c),). In Fig. 2 (b) we see that CN2, which is commonly altered to achieve calibration, illustrates a similar situation where a 10% decrease from the initial value provides a wide range of desirable solutions. The implications of wide ranges in CN2 providing desirable solutions are discussed in Sect. 3.3. Many good solutions (medium to dark blue regions) exist within the parameter ranges evaluated, and each is associated with a substantially different representation of the physical system. Local maximas can be found throughout the parameter space and the automated routine could potentially pick any one of the many solutions that are deemed good based on a lumped metric. Furthermore, selection of the parameter values is often conditioned on a specified stopping criteria, which may terminate calibration prematurely once an acceptable NSE value is obtained. The same three plots were generated

with R^2 and PBIAS as the performance metric and are shown in the SI (Fig. S3 and Fig. S4). These plots indicate that multi objective criteria may not fully resolve the challenges with equifinality.

3.3 Parameter selection based on sensitivity

Whether automated or manual calibration is used, a common approach is to adjust the parameters that display the highest sensitivity (Madsen, 2003;Hill, 1998). This approach is attractive because improvements in calibration can be achieved with a minimal number of adjustments. However, sensitivity alone should not be the criterion for parameter set selection for calibration because the most sensitive parameter is not necessarily the one causing the divergence between the modeled and measured values (See Fig. 2 (c) above, where calibration can be achieved through many different parameter combinations). For example, the CN2 is a highly sensitive parameter and thus minor adjustments can drastically alter the hydrograph. For this reason, CN2 is commonly used as a 'sledge hammer' in calibration to make significant changes across a wide range of flow values in an effort to achieve a higher performance metric value, regardless of whether or not there is reason to believe modified CN2 are justifiable.

To illustrate this behavior, we varied CN2 ±10% and show the corresponding stream flow outcome for 1000 runs for the South Branch gage in the RRB. Varying CN2 by ±10% is a common practice to achieve calibration (Williams et al., 2011). We can see over 50% variability in peak flows for a ±10% adjustment in CN2 (Fig. 3 a-c). The implications of such changes when evaluating scenarios could render the model incapable of capturing the effects of certain management practices, as most practices alter CN ± 5%. To demonstrate this point, we altered CN2 ± 5% from the commonly employed initialized value based on local land use and management. In Fig. 3 (a-c), we can see the lower and upper bounds that represents land use change or a management practice are completely within the calibration band. The conclusion that can be drawn from Fig. 3 (a-c) is that certain management scenarios such as land use or management change can be within the realm of uncertainty. There are no differences in the upper limits of the +10% CN2 and +5% CN2 bands. The dominance of karst features consisting of sinkholes and sinking streams renders the highly sensitive CN2 insensitive beyond a certain threshold and does not translate to increased peak flows. Any increase in runoff from increased CN2 is damped by transmission losses in the tributary and main channels. However, the lower limits for the -10% CN2 case extends beyond the -5% CN2 case. This indicates an upper threshold and not a lower threshold for the ranges explored.

3.4 Choosing parameters through physics

SWAT and many other hydrologic models are physically based, where theoretically, we should be able to directly link changes in hydrologic parameters to changes in predicted streamflow outcomes. However, our ability to link parameters and streamflow outcomes quickly becomes difficult when the parameter used to explain the outcome is nested within mathematical structures that contain multiple other parameters. It becomes even more challenging when these parameters are derived primarily through calibration and have interdependences. Notwithstanding these limitations and problems associated with parameter interactions and threshold based sensitivities, manually or automatically adjusting parameters within a physically meaningful range can

serve as an effective means for improving model calibration. We show how physics can be used to understand model behavior in the time domain and in the frequency domain.

3.4.1 Parameter behavior in the time domain

We illustrate challenges in linking parameters and streamflow outcomes using a basin level SWAT parameter called snow pack temperature lag factor, also referred as TIMP and the threshold temperature for snowmelt, referred to as SMTMP. Both these parameters are pure calibration parameters (Wang and Melesse, 2005) and are discussed as such in this paper. Several SWAT parameters are specified at the spatial resolution of the basin, e.g., TIMP and SMTMP. Adjustment of these parameters will have basin wide implications. Within SWAT, TIMP is used as a proxy for snow pack density, snow pack depth, exposure and other factors that affect the temperature of the snow pack. A higher value (max= 1) indicates a strong influence of current day temperature on the temperature of the snow pack and a low value (min = 0.01) indicates a greater influence of the previous day's temperature. One possible way to select parameters could be by identifying factors and periods when they will have an effect. By comparing with a measured parameter like air temperature, we can identify candidates for calibration provided they are sensitive for the time period of interest. This approach goes beyond a blanket sensitivity analysis which is explicitly blind to the physics. The grey box in Fig. 4 (a) and (b) highlights time periods with significant offsets between the two runs, demonstrating that the parameter is sensitive when the air temperature persists in sub-zero temperatures. Same is the case for Fig. 4 (d). When temperature rises above zero, the parameter has no effect in the outcome (See Fig. 4 (c) inset). Scrutinizing the effects of TIMP in this way allows us to better understand its influence on the magnitude as well as timing of streamflow.

3.4.2 Parameter behavior in the frequency domain (via wavelet analysis)

Time domain analysis such as presented in Sect. 3.4.1, allows us to see changes in streamflow response when a parameter is adjusted visually. Frequency domain analysis is a complementary approach to evaluate the effects of parameter adjustment. Analyzing the frequency content of the signal using wavelets can provide the frequency content of the two signals that is localized in time (Addison, 2005). The wavelet approach offers a targeted form of sensitivity analysis compared to the blanket approach using only lumped metrics.

The wavelet transform is an integral transform that allows for the evaluation of the temporal evolution of frequency content contained in a nonstationary signal (Pichot et al., 1999;Foufoula-Georgiou and Kumar, 2014) such as streamflow. Wavelets are functions that have a specified window where they are non-zero. These functions can be dilated or stretched to check the presence of a particular frequency in the signal. In wavelet analysis, lower scales correspond to higher frequencies and vice versa. The resulting wavelet coefficients represent the correlation between the signal at different points along the signal and at multiple scales (Percival, 1995). At any scale along the vertical axis and at any point representing time along the horizontal axis, larger coefficients are indicative of greater correspondence between the analyzing wavelet and the analyzed signal (e.g., streamflow). Each scale corresponds to the width of the wavelet or the length of the signal that will be analyzed.

Following a wavelet transform, comparison in the frequency domain can be thought of as a way to compare the constituent shapes (individual frequencies) present in each signal. The coefficients that result from each signal can be compared to each other via wavelet coherence, which is analogous to the R^2 in statistics (Liu, 1994). A coefficient value close to one indicates the frequency/shape is present in both signals for the time period under consideration. Wavelet coherence analysis of the pre-parameter adjustment signal and the post-parameter adjustment signal allows us to identify specific scales and time periods when parameter adjustments will impact the signal of interest (here, streamflow). The Sect. S1 in the SI contains a brief summary on wavelets, wavelet transform and wavelet coherence to keep the paper self-contained, however, see (Daubechies, 1992) or (Foufoula-Georgiou and Kumar, 2014) and the references therein.

Figure 5(a-d) shows wavelet coherence between pre- and post-adjustment signals for four pure calibration parameters. Hotter colors (red) indicate the shape is absent in either one of the signals and they don't match, whereas cooler colors (blue) indicate that the shapes match between the two signals. The extent (time and scale) and influence (hotter colors) of a parameter is coded in color. We can see the scales and time periods where the parameters are sensitive. For example, (See Fig. 5 (a)) is only sensitive up to approximately the 32-day band and between November and March. So, if there are mismatches between measured and simulated flow outside this time period or beyond a 32-day band, TIMP cannot resolve the mismatch. Similarly, for SFTMP, SMFMX and SMFMN and CN_FROZ, shown in Fig. 5 (b-d), adjustment only affects some portions of the signal and only some scales. In Fig. 5 (d), the time periods where there are mismatches are not continuous, indicating selective bands of influence or sensitivity. The tools developed to conduct this analysis are distributed as part of the HydroME Toolbox.

3.5 Parameter classification

To address challenges with parameter choice, whether using manual or fully automated procedures and in the time or frequency domain, we emphasize that the modeler consider different classes of parameters. Our proposed approach attempts to choose the most relevant and parsimonious set of parameters to guide calibration, minimize physical distortion of the hydrologic system being modeled, focus calibration and validation on relevant hydrologic metrics, and reduce the problem of equifinality. We differentiate calibration parameters into three categories, namely pure, derived and measured. Considering parameters within these three categories provides a rational framework to move beyond calibration primarily based on sensitivity analysis. In this hierarchical approach for parameter adjustment, we prioritize parameters sets in order of least certain (pure) to most certain (measured). We define the three parameter sets and discuss assumptions and caveats of this approach in the following paragraphs.

3.5.1 Pure calibration parameters (Stage 1)

Parameters that have no measured or derived basis and the values for which are commonly determined exclusively through calibration are defined in this paper as pure calibration parameters. These parameters are referred by some authors as artifacts of model structure and cannot be measured in the field (Singh and Frevert, 2002). For example, the TIMP parameter used in SWAT serves as a proxy for snow pack density, snow pack depth, exposure and other factors that affect the temperature of the

snow pack (Neitsch et al., 2011). This parameter cannot be directly measured in the field. Adjustment of pure calibration parameters is typically based on its effect on a performance metric, guided perhaps by a hydrologic modeler's intuition, but with little or no physical basis. Typically, adjustment of these parameters is based on literature reported values for the area. However, the robustness of these literature reported values is difficult to assess, especially when scale, thresholds, site specific considerations, and other challenges described in Singh and Frevert (2002) might have influenced their values.

Pure calibration parameters come with the highest level of uncertainty among all the SWAT parameters. In SWAT, these parameters are specified with default values during the model building step. Therefore, we propose that adjustment of this parameter set should be explored first to achieve calibration because their default values are supported by the least actual (measured or derived) information. Figure 6 shows the parameters that were considered 'pure calibration parameters' for the two study watersheds.

3.5.2 Derived parameters (Stage 2)

Parameters derived from measurements or observations either through relationships, calculations or lookup tables are defined as derived parameters. They theoretically contain not only the uncertainties and errors associated with the measured values on which they are based, but also uncertainties and errors due to their specification. These parameters are often supported by extensive surveys or research that has been generalized in a way that is well accepted within the community, if not without uncertainties. For example, CN2 is commonly adjusted in SWAT to reduce mismatch between modeled and observed streamflow. Even though specification of CN2 can be subjective to certain extent, the variability of these values have consensus in the literature and have a sound scientific basis. Therefore we can be more certain about them than pure calibration parameters (Dahlke et al., 2012;Steenhuis et al., 1995). Other examples of derived parameters include Manning's roughness coefficients for tributary and main channels (CH_N1 and CH_N2) and others shown in Fig. 6.

3.5.3 Measured parameters (Stage 3)

The third set, referred to in this paper as 'measured parameters,' are those directly based on physical measurements and theoretically will have the least uncertainty or error among the three sets. This list includes all parameters that can be directly measured in the field. We propose that they should be adjusted only in cases where a rationale can be provided. Examples of these parameters include soil hydraulic conductivity and soil texture. Physically meaningful ranges could be specified during automatic calibration during stage 3 to constrain uncertainty bands associated with these values, due to measurement errors or data aggregation errors.

3.5.4 Flexible nature of parameter classification

Depending on the watershed of interest and available data, the set of pure calibration parameters may be expanded to include parameters that might otherwise be considered derived or measured, but are insufficiently constrained in your particular setting. For example, watersheds throughout Minnesota are artificially drained by perforated sub-surface tubing, referred to as 'tile

drainage' but, the distribution or density of these sub-surface tiles are seldom known. In such cases, all parameters associated with tile drainage become either pure or derived calibration parameters (e.g., tile spacing, depth to tile and tile diameter) depending on the amount of information that is available. Sensible initial values for these parameters should be based on sound engineering practices as described in engineering design manuals such as USDA-NRCS (1971). However, their actual depth, spacing, size and orientation is decided by individual farmers and is contingent on localized factors such as micro-topography and slight differences in soil properties, as well as farmer experience and financial limitations. As their location and distribution are unknown, we propose that the tile parameters should be treated as a pure calibration parameter or perhaps derived parameters in cases where personal communications with farmers or direct observations inform our estimates. Figure 6 illustrates the flexible nature of the calibration parameters.

4 Meaningful measures for evaluating model performance

Models can be developed for a variety of applications and the metrics used to evaluate them should be targeted to the application of interest. For example, it is generally recognized that a good agreement between modeled and measured streamflow should involve: (1) shape of the hydrograph, (2) the timing, rate and the volume of peak flows, and (3) low flows as described in Madsen (2000). Others have reported five hydrologic metrics that are relevant for ecology and water quality, including; (1) magnitude, (2) frequency, (3) duration, (4) timing and (5) rate of hydrograph rise and fall (e.g., (Liu et al., 2011;Zolezzi et al., 2009;Mathews and Richter, 2007;Lytle and Poff, 2004). Yet, in practice metrics used to evaluate their performance are commonly limited to visual inspection, RMSE, R^2 or NSE. Better targeted model performance metrics facilitate more efficient and robust calibration and will help the modeller make a more compelling case that the model is suitable for the application of interest.

Visual hydrograph inspection continues to be the most widely used technique for hydrologic model evaluation (Seibert et al., 2016;Reusser et al., 2009;Ehret and Zehe, 2011). Dividing the hydrograph into distinct time periods and adjusting parameters by trial and error is still commonly done during model calibration (Singh and Frevert, 2002). However, objective and quantitative metrics are preferable when reporting results. In the following sub-sections we introduce simple variations in how common classical error measures are reported and show that by combing them with other graphical representations to evaluate model performance they can serve as a powerful model evaluation tool. The HydroME Toolbox (macro-enabled Microsoft Excel workbook) that generates the graphs and performance metrics described below can be downloaded for free from: https://qcnr.usu.edu/labs/belmont_lab/resources. Specifically, we show that combining lumped error measures with other graphical representations of measured and model simulated flows, we can increase efficiency and targeting of specific flow conditions during model calibration and evaluation.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-121, 2017
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 6 March 2017

4.1 Lumped metrics

Automated and intervention type calibration in hydrology generally rely on a lumped model performance metric such as NSE (a normalized measure ranging between – inf to 1.0) or R^2 (Krause et al., 2005;Reusser et al., 2009). Such metrics are useful insofar as they are objective measures that can be used to quantify model behavior and can be readily and quantitatively compared among models (Krause et al., 2005). Studies have highlighted the need for other measures in a multi-objective criteria approach for a robust evaluation of model performance (Yapo et al., 1998;Singh and Frevert, 2002). The use of a single measure serves well in an automatic calibration context, but does not necessarily capture the most relevant aspects of model performance (Reusser et al., 2009). For example, the quadratic formulation of NSE or R^2 emphasizes higher magnitude flows, relative to lower flows, the latter of which may be critical for many ecological and water quality applications (Criss and Winston, 2008). Furthermore, the character of a watershed and the time step of evaluation have been shown to influence evaluation metrics and could potentially place false confidence in the ability of a hydrologic model to simulate streamflow (Schaefli and Gupta, 2007;Belmont et al., 2016b;Krause et al., 2005). These variance-based lumped error measures also do not separate different flow components or time dependencies (Liu et al., 2011).

We advocate for the continued use of lumped model performance metrics as they serve a useful purpose to determine offsets in flow magnitude (See Table 1). Nevertheless, given the current level of sophistication of hydrologic modeling and increasing demand for models that target specific hydrologic metrics (e.g., summer base flows, timing and rate of the rising limb of the snowmelt hydrograph), we propose that the community is poised to move beyond these basic metrics.

Windowing the signal into different time periods (seasons, months etc.) can provide useful information regarding model performance. Here we propose to go a step further and use box plots to report performance measures such as NSE and R^2 by aggregating them into annual, seasonal or monthly intervals as shown in Fig. 7. The advantage of such a representation is that these plots can clearly show the range and distribution of the performance metric as opposed to just a central tendency. Generally, the mean of a performance metric is reported, but that approach may obscure how well the model is performing for the time period of interest. For example, Fig. 7 shows that model performance varies considerably during the fall and winter months and is captured by both metrics; NSE and R^2. Additional useful information is the Interquartile Range (IQR), which is lowest for spring irrespective of the metric considered and shows clustering around the central tendency. Additionally, the negative skew shows that most of the time the model performance is not as good as what is represented by just the central tendency.

4.2 Euclidean distance, Empirical Q-Q plots and FDCs

Rather than simple visual comparison on modeled versus measured hydrographs, the HydroME Toolbox plots the difference between the two (Q measured – Q predicted), which makes it easier to identify large or systematic deviations. In addition, the empirical Quantile-Quantile plot is a graphical technique that can be used to roughly determine how well measured flow is predicted by the model (Helsel and Hirsch, 1992). An example from the Le Sueur SWAT model (Little Cobb gage for 1996 to

2000) is shown in Fig. 6 to illustrate the usefulness of these simple techniques. Figure 7(a) shows mostly positive deviations, indicating the model is under-predicting streamflow. Although the reason cannot be determined from this plot alone, the plot illustrates the frequency, timing and magnitude of offsets, which led us to suspect that drain tile connectivity could be the cause of model under-prediction. The Empirical Quantile-Quantile (EQQ) plot is a plot between the quantiles of the measured flow and model predicted flow for the same time period. In the EQQ plot in Fig. 7(b), departure from the 1:1 reference line is minimal, with only slight biases at very low and high flows. Considering both of these plots, parameter adjustment during calibration can target those times and flows that are observed to be skewed. For the purpose of reporting model performance, the empirical Q-Q plot can provide information that cannot be gleaned from a lumped metric alone.

Flow Duration Curves (FDCs) are simple, yet powerful and insightful graphical representations that allow for the evaluation of a full range of flows that occur during a period of interest. FDCs are used extensively in hydrologic analyses and have been qualitatively used for evaluating model performance (Son and Sivapalan, 2007), but are rarely used during calibration of a hydrologic model. While flow duration curves eliminate the timing of flows, they are a quick and easy means to evaluate which components of the flow regime (high, moderate, or low flows) are systematically under- or over-predicting. Figure 7 (c) illustrates the calibrated model is capable to capturing most flows below 80% exceedance. The uncalibrated model deviates from measured flows across the full range of flows.

5 Reporting: Model performance metrics

In addition to the meaningful measures described in the previous section, where model evaluation was performed in the time domain, frequency domain analysis can serve as a powerful tool for model evaluation. Watershed behavior is complex and representing the data in both time and frequency domains can allow us to glean information that may be obscured when examined solely in either domain. In this paper we introduce two frequency domain methods that can evaluate the shape of the hydrograph; namely, magnitude squared coherence and wavelet coherence.

5.1 Model evaluation using magnitude squared coherence

We can evaluate the performance of a model by determining if the frequency components of the measured signal are present in the simulated signal. The magnitude squared coherence approach is used to compare the frequency content of measured and simulated signals and provides a measure of their similarity (Ropella et al., 1989). The technique captures all the frequencies present in the analyzed signal without localizing the frequency. The coherence estimate provides a general sense of model performance and can reveal which flow frequencies in the measured signal were captured by the model. The specific coherence technique used here is called the magnitude squared coherence via weighted overlapped segment averaging. It measures the strength of association between two stationary stochastic processes (Wang et al., 2004). With the assumption of a stationary stochastic process, we ignore the time at which the frequency occurred in the signal. We acknowledge that such an assumption could result in a false positive (i.e., a given frequency could be present in both signals, but occur at different times), however,

this issue can be resolved with the wavelet analysis, which has a different set of advantages and limitations, as discussed in Sect. 5.2. When the processes that determine the flow outcome are independent or absent, a coherence of zero should be expected and a value of one when they are dependent or present. Prior to calibration, most frequencies that characterize the observed streamflow are absent in the uncalibrated model-predicted signal (Fig. 10). After calibration, the frequency spectra of observed flows and model predictions are far more similar. For example, note the higher coherence values across the spectrum and especially for frequencies below 0.25 (wavelength of 4 days) in Fig. 10). It can be seen even with calibration; some frequencies are absent in the simulated signal. The magnitude-squared coherence, $\gamma_{xy}^2(f)$ is defined as shown in Eq. (1),

$$\gamma_{xy}^2(f) = \frac{|P_{xy}(f)|^2}{P_{xx}(f)P_{yy}(f)} \quad (1)$$

where, $P_{xy}(f)$ is a complex cross spectral density, $P_{xx}(f)$ and $P_{yy}(f)$ are auto spectral densities at a particular frequency f (Wang et al., 2004).

5.2 Wavelet coherence for model performance evaluation

Comparing model performance using the magnitude squared coherence approach compares all frequencies present in the signal, however it cannot localize the mismatch in time. This results in the loss of time information (Si, 2008). Moreover, the global nature of lumped error metrics (e.g., NSE, R^2) is of little use when event scale predictions or time resolution of model performance is needed. Instead, coherence calculated using wavelets can provide this information. Specifically, the wavelet time-frequency domain representation of measured and simulated signals allows us to clearly see (1) if there are any patterns in the model's inability to simulate observed flows, and (2) the times and timescales for which the model was able to capture the measured streamflow. In Fig. 10 (a) and (c), the residuals show the time periods when there are differences between the measured and predicted streamflow. This time domain view can provide the modeler insight into the time and magnitude of discrepancies. Similar to the wavelet coherence estimate between the pre- and post-adjustment signals as discussed in Sect. 3.4.2, a value of one indicates the shapes in both the measured and predicted signals are the same. In Fig. 10 (b), the wavelet coherence plot shows that frequencies below the 32-day band are generally not captured well, except for a few locations. Beyond the 32-day band, we see a nearly 100% capture of the measured signal by the model. In Fig. 10 (d), we see mismatches up to the 128-day band during certain time periods. However, there are time periods when the higher frequencies are captured well by the model.

6 Conclusions and Discussions

Watershed hydrology models have advanced considerably over the past decade. With increasing capabilities and demands on such models, it is essential that the community embrace more targeted and meaningful metrics for model evaluation. In this paper we have introduced and tested a suite of new tools for model evaluation, suggest a hierarchical approach to the selection of parameters to adjust during calibration, and discuss the benefits of evaluating models in both time and frequency domains.

Methods described in this paper were applied to two model instances representing physical environments that challenge calibration in different ways.

Most hydrology models require some form of calibration as many of the parameters are only conceptual and cannot be measured in the field. Even parameters that can be measured in the field may not always be available. Thus, a hierarchical approach to parameter adjustment, starting with the most uncertain parameters, is recommended. We show equifinality and interdependencies among parameters present formidable challenges for fully automated calibration procedures. For example, automated routines depend on lumped metrics and cannot uniquely identify parameter sets that optimize calibration while minimizing distortion of the physical system, even with the use of multi-objective lumped metrics. Many parameter combinations, representing different characterizations of the physical system, result in local maxima for numerous evaluation criteria, as shown in Fig. 2 and associated figures in SI. Further, there are pitfalls of using a blanket sensitivity-based approach in the selection of parameters to adjust for calibration, which has the potential to render a model unable to reliably simulate certain management practices. Reliance on the Curve Number method widely within the SWAT framework to inform management decisions for implementing various conservation practices must be implemented with caution.

We illustrate choosing parameters through both time and frequency domains have complementary strengths that can be leveraged to identify parameters for calibration. Scenarios can be evaluated by comparing different values of the same parameter along with related observations to identify time periods and magnitudes of mismatch. Frequency domain analysis can provide insight in cases when there are complex interactions between parameters as physics-based explanations for parameter adjustments may be inadequate. Wavelet transform provides the description of frequencies in a hydrologic signal that are localized by time. The magnitude squared coherence and wavelet analyses are most useful in the latter stages of calibration and model evaluation, after major parameter adjustments have been made based on lumped metrics and other time domain metrics.

We provide a guideline for prioritizing parameters during calibration based on the information we have about the particular parameter. The primary motivation for such a classification is to minimize physical distortion through parameter adjustment. In this paper we classify parameters into three categories, namely; pure, derived and measured based on the amount of uncertainty. We also demonstrate the flexible nature of the classification as amount information known for any particular parameter can vary with study area or access to information. Modifying model structure can facilitate calibration. However, we encourage using a model structure that is appropriate for time scales of evaluation, the questions of interest, and the environment under consideration.

Lumped metrics make it easy to evaluate model performance with one all-inclusive value. However, reporting lumped metrics by segregating them annually or monthly using box-plots can reveal considerable information about model performance that are otherwise obscured by the central tendency based approach. Euclidian distance measure can be used to identify magnitude mismatches that are localized in time. Empirical Q-Q plots show the quantiles that were captured by the model and aid in calibration. FDCs can be another technique that can reveal what the model is able to simulate. Further exploration is needed for specifically identifying parameters that can help remove any mismatch.

Frequency and time-frequency domain measures were used to report model performance. Magnitude squared coherence was used to compare the measured and predicted signals, which highlighted that even for a NSE value of 0.74 many of the frequencies were missing in the predicted flow. This approach, however, did not resolve the time of mismatch. The use of wavelet coherence allowed for a time resolution of frequency comparison. The measured and simulated flows showed a general mismatch in the higher frequencies which were delineated both by time and scale. Like lumped metrics or other time domain model performance measures, we cannot fully and unambiguously separate the sources of mismatch, which could be attributed to data, model structure or the calibration itself.

7 Acknowledgements

This work was made possible with support from the National Science Foundation (NSF ENG 1209445) grant.

References

Addison, P. S.: Wavelet transforms and the ECG: a review, Physiological Measurement, 26, R155, 2005.

Baffaut, C., and Benson, V.: Modeling flow and pollutant transport in a karst watershed with SWAT, Trans. Asabe, 52, 469-479, 2009.

Belmont, P.: Floodplain width adjustments in response to rapid base level fall and knickpoint migration, Geomorphology, 128, 92-102, http://dx.doi.org/10.1016/j.geomorph.2010.12.026, 2011.

Belmont, P., Gran, K. B., Schottler, S. P., Wilcock, P. R., Day, S. S., Jennings, C., Lauer, J. W., Viparelli, E., Willenbring, J. K., Engstrom, D. R., and Parker, G.: Large Shift in Source of Fine Sediment in the Upper Mississippi River, Environmental Science & Technology, 45, 8804-8810, 10.1021/es2019109, 2011.

Belmont, P., Willenbring, J. K., Schottler, S. P., Marquard, J., Kumarasamy, K., and Hemmis, J. M.: Toward generalizable sediment fingerprinting with tracers that are conservative and nonconservative over sediment routing timescales, Journal of Soils and Sediments, 14, 1479-1492, 10.1007/s11368-014-0913-5, 2014.

Belmont, P., Dogwiler, T., and Kumarasamy, K.: An integrated sediment budget for the Root River watershed, Southeastern Minnesota. Final Report to the Minnesota Department of Agriculture., 2016a.

Belmont, P., Stevens, J. R., Czuba, J. A., Kumarasamy, K., and Kelly, S. A.: Comment on "Climate and agricultural land use change impacts on streamflow in the upper midwestern United States," by Satish C. Gupta et al, Water Resources Research, 52, 7523-7528, 10.1002/2015WR018476, 2016b.

Beven, K., and Freer, J.: Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, Journal of Hydrology, 249, 11-29, http://dx.doi.org/10.1016/S0022-1694(01)00421-8, 2001.

Beven, K.: A manifesto for the equifinality thesis, Journal of Hydrology, 320, 18-36, http://dx.doi.org/10.1016/j.jhydrol.2005.07.007, 2006.

Beven, K. J.: Rainfall-runoff modelling: the primer, John Wiley & Sons, 2011.

Butts, M. B., Payne, J. T., Kristensen, M., and Madsen, H.: An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation, Journal of Hydrology, 298, 242-266, http://dx.doi.org/10.1016/j.jhydrol.2004.03.042, 2004.

SWAT literature database for peer-reviewed journal articles: https://www.card.iastate.edu/swat_articles/, access: 12/12/2016, 2016.

Chen, J., and Wu, Y.: Advancing representation of hydrologic processes in the Soil and Water Assessment Tool (SWAT) through integration of the TOPographic MODEL (TOPMODEL) features, Journal of Hydrology, 420–421, 319-328, http://dx.doi.org/10.1016/j.jhydrol.2011.12.022, 2012.

Chow, T. V.: Open channel hydraulics, McGraw-Hill Book Company, Inc; New York, 1959.

Criss, R. E., and Winston, W. E.: Do Nash values have value? Discussion and alternate proposals, Hydrological Processes, 22, 2723-2725, 10.1002/hyp.7072, 2008.

Dahlke, H. E., Easton, Z. M., Walter, M. T., and Steenhuis, T. S.: Field Test of the Variable Source Area Interpretation of the Curve Number Rainfall-Runoff Equation, Journal of Irrigation and Drainage Engineering, 138, doi:10.1061/(ASCE)IR.1943-4774.0000380, 2012.

Daubechies, I.: Ten lectures on wavelets, SIAM, 1992.

Efstratiadis, A., and Koutsoyiannis, D.: One decade of multi-objective calibration approaches in hydrological modelling: A review, Hydrological Sciences Journal, 55, 58-78, 10.1080/02626660903526292, 2010.

Ehret, U., and Zehe, E.: Series distance – an intuitive metric to quantify hydrograph similarity in terms of occurrence, amplitude and timing of hydrological events, Hydrol. Earth Syst. Sci., 15, 877-896, 10.5194/hess-15-877-2011, 2011.

Fisher, T. G.: Chronology of glacial Lake Agassiz meltwater routed to the Gulf of Mexico, Quaternary Research, 59, 271-276, http://dx.doi.org/10.1016/S0033-5894(03)00011-5, 2003.

Foufoula-Georgiou, E., and Kumar, P.: Wavelets in geophysics, Academic Press, 2014.

Gassman, P. W., Reyes, M. R., Green, C. H., and Arnold, J. G.: The Soil and Water Assessment Tool: Historical Development, Applications, and Future Research Directions, 50, 10.13031/2013.23637, 2007.

Gran, K. B., Belmont, P., Day, S. S., Finnegan, N., Jennings, C., Lauer, J. W., and Wilcock, P. R.: Landscape evolution in south-central Minnesota and the role of geomorphic history on modern erosional processes, GSA Today, 21, 7-9, 2011.

Gran, K. B., Finnegan, N., Johnson, A. L., Belmont, P., Wittkop, C., and Rittenour, T.: Landscape evolution, valley excavation, and terrace development following abrupt postglacial base-level fall, Geological Society of America Bulletin, 125, 1851-1864, 10.1130/b30772.1, 2013.

Gupta, H. V., Sorooshian, S., and Yapo, P. O.: Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information, Water Resources Research, 34, 751-763, 10.1029/97WR03495, 1998.

Helsel, D. R., and Hirsch, R. M.: Statistical methods in water resources, Elsevier, 1992.

Hill, M. C.: Methods and guidelines for effective model calibration; with application to UCODE, a computer code for universal inverse modeling, and MODFLOWP, a computer code for inverse modeling with MODFLOW, Report 98-4005, 1998.

Jennings, C.: Draft Digital Reconnaissance Surficial Geology and Geomorphology of the Le Sueur River Watershed (Blue Earth, Waseca, Faribault, and Freeborn Counties in South-Central MN), Map scale, 1, 000, 2010.

Knox, J. C.: Historical Valley Floor Sedimentation in the Upper Mississippi Valley, Annals of the Association of American Geographers, 77, 224-244, 10.1111/j.1467-8306.1987.tb00155.x, 1987.

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

Liu, P. C.: Wavelet spectrum analysis and ocean wind waves, Wavelets in geophysics, 4, 151-166, 1994.

Liu, Y., Brown, J., Demargne, J., and Seo, D.-J.: A wavelet-based approach to assessing timing errors in hydrologic predictions, Journal of Hydrology, 397, 210-224, http://dx.doi.org/10.1016/j.jhydrol.2010.11.040, 2011.

Lytle, D. A., and Poff, N. L.: Adaptation to natural flow regimes, Trends in Ecology & Evolution, 19, 94-100, http://dx.doi.org/10.1016/j.tree.2003.10.002, 2004.

Madsen, H.: Automatic calibration of a conceptual rainfall–runoff model using multiple objectives, Journal of Hydrology, 235, 276-288, http://dx.doi.org/10.1016/S0022-1694(00)00279-1, 2000.

Madsen, H., Wilson, G., and Ammentorp, H. C.: Comparison of different automated strategies for calibration of rainfall-runoff models, Journal of Hydrology, 261, 48-59, http://dx.doi.org/10.1016/S0022-1694(01)00619-9, 2002.

Madsen, H.: Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives, Advances in Water Resources, 26, 205-216, http://dx.doi.org/10.1016/S0309-1708(02)00092-1, 2003.

Mathews, R., and Richter, B. D.: Application of the Indicators of Hydrologic Alteration Software in Environmental Flow Setting1, JAWRA Journal of the American Water Resources Association, 43, 1400-1413, 10.1111/j.1752-1688.2007.00099.x, 2007.

MPCA: Root River Watershed Monitoring and Assessment Report 2012.

Neitsch, S. L., Williams, J., Arnold, J., and Kiniry, J.: Soil and water assessment tool theoretical documentation version 2009, Texas Water Resources Institute, 2011.

Percival, D. P.: On estimation of the wavelet variance, Biometrika, 82, 619-631, 10.1093/biomet/82.3.619, 1995.

Pichot, V., Gaspoz, J.-M., Molliex, S., Antoniadis, A., Busso, T., Roche, F., Costes, F., Quintin, L., Lacour, J.-R., and Barthélémy, J.-C.: Wavelet transform to quantify heart rate variability and to assess its instantaneous changes, Journal of Applied Physiology, 86, 1081-1091, 1999.

Reusser, D. E., Blume, T., Schaefli, B., and Zehe, E.: Analysing the temporal dynamics of model performance for hydrological models, Hydrology And Earth System Sciences, 13, 999-1018, 10.5194/hess-13-999-2009, 2009.

Ropella, K. M., Sahakian, A. V., Baerman, J. M., and Swiryn, S.: The coherence spectrum. A quantitative discriminator of fibrillatory and nonfibrillatory cardiac rhythms, Circulation, 80, 112-119, 10.1161/01.cir.80.1.112, 1989.

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., Behringer, D., Hou, Y.-T., Chuang, H.-y., Iredell, M., Ek, M., Meng, J., Yang, R., Mendez, M. P., Dool, H. v. d., Zhang, Q., Wang, W., Chen, M., and Becker, E.: The NCEP Climate Forecast System Version 2, Journal of Climate, 27, 2185-2208, 10.1175/jcli-d-12-00823.1, 2014.

Sanborn, S. C., and Bledsoe, B. P.: Predicting streamflow regime metrics for ungauged streamsin Colorado, Washington, and Oregon, Journal of Hydrology, 325, 241-261, http://dx.doi.org/10.1016/j.jhydrol.2005.10.018, 2006.

Santhi, C., Arnold, J. G., White, M., Di Luzio, M., Kannan, N., Norfleet, L., Atwood, J., Kellogg, R., Wang, X., Williams, J. R., and Gerik, T.: Effects of Agricultural Conservation Practices on N Loads in the Mississippi–Atchafalaya River Basin, Journal of Environmental Quality, 43, 1903-1915, 10.2134/jeq2013.10.0403, 2014.

Schaefli, B., and Gupta, H. V.: Do Nash values have value?, Hydrological Processes, 21, 2075-2080, 10.1002/hyp.6825, 2007.

Schottler, S. P., Ulrich, J., Belmont, P., Moore, R., Lauer, J. W., Engstrom, D. R., and Almendinger, J. E.: Twentieth century agricultural drainage creates more erosive rivers, Hydrological Processes, 28, 1951-1961, 10.1002/hyp.9738, 2014.

Seibert, S. P., Ehret, U., and Zehe, E.: Disentangling timing and amplitude errors in streamflow simulations, Hydrol. Earth Syst. Sci., 20, 3745-3763, 10.5194/hess-20-3745-2016, 2016.

Si, B. C.: Spatial Scaling Analyses of Soil Physical Properties: A Review of Spectral and Wavelet Methods All rights reserved. No part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher, Vadose Zone Journal, 7, 547-562, 10.2136/vzj2007.0040, 2008.

Singh, V. P., and Frevert, D. K.: Mathematical models of large watershed hydrology, Water Resources Publication, 2002.

Son, K., and Sivapalan, M.: Improving model structure and reducing parameter uncertainty in conceptual water balance models through the use of auxiliary data, Water Resources Research, 43, n/a-n/a, 10.1029/2006WR005032, 2007.

Sorooshian, S., and Gupta, V. K.: Automatic calibration of conceptual rainfall-runoff models: The question of parameter observability and uniqueness, Water Resources Research, 19, 260-268, 10.1029/WR019i001p00260, 1983.

Steenhuis, T. S., Winchell, M., Rossing, J., and Zollweg, J. A.: SCS Runoff Equation Revisited for Variable-Source Runoff Areas, Journal of Irrigation and Drainage Engineering, 121, doi:10.1061/(ASCE)0733-9437(1995)121:3(234), 1995.

Stout, J. C., Belmont, P., Schottler, S. P., and Willenbring, J. K.: Identifying Sediment Sources and Sinks in the Root River, Southeastern Minnesota, Annals of the Association of American Geographers, 104, 20-39, 10.1080/00045608.2013.843434, 2014.

Thorleifson, L.: Review of Lake Agassiz history, Sedimentology, Geomorphology, and History of the Central Lake Agassiz Basin. Geological Association of Canada Field Trip Guidebook B, 2, 55-84, 1996.

Tolson, B. A., and Shoemaker, C. A.: Dynamically dimensioned search algorithm for computationally efficient watershed model calibration, Water Resources Research, 43, 10.1029/2005WR004723, 2007.

Troelstrup, N., and Perry, J. A.: Water quality in southeastern Minnesota streams: observations along a gradient of land use and geology, Journal of the Minnesota Academy of Science, 55, 6-13, 1989.

Wang, S.-Y., Liu, X., Yianni, J., Christopher Miall, R., Aziz, T. Z., and Stein, J. F.: Optimising coherence estimation to assess the functional correlation of tremor-related activity between the subthalamic nucleus and the forearm muscles, Journal of Neuroscience Methods, 136, 197-205, http://dx.doi.org/10.1016/j.jneumeth.2004.01.008, 2004.

Wang, X., and Melesse, A.: Evaluation of the SWAT model's snowmelt hydrology in a northwestern Minnesota watershed, Transactions of the ASAE, 48, 1359-1376, 2005.

Weiguo, H., Zhengwei, Y., Liping, D., and Peng, Y.: A Geospatial Web Service Approach for Creating On-Demand Cropland Data Layer Thematic Maps, 57, 10.13031/trans.57.10020, 2014.

Weiler, M., and Beven, K.: Do we need a Community Hydrological Model?, Water Resources Research, 51, 7777-7784, 10.1002/2014WR016731, 2015.

Wenger, S. J., Luce, C. H., Hamlet, A. F., Isaak, D. J., and Neville, H. M.: Macroscale hydrologic modeling of ecologically relevant flow metrics, Water Resources Research, 46, n/a-n/a, 10.1029/2009WR008839, 2010.

Wilcock, P. R., and Belmont, P.: Identifying sediment sources in the Minnesota River Basin, 2009.

Williams, J., Kannan, N., Wang, X., Santhi, C., and Arnold, J.: Evolution of the SCS runoff curve number method and its application to continuous runoff simulation, Journal of Hydrologic Engineering, 17, 1221-1229, 2011.

WRC, and MPCA: State of the Minnesota River. Summary of surface water quality monitoring 2000-2008, 42, 2009.

Yapo, P. O., Gupta, H. V., and Sorooshian, S.: Multi-objective global optimization for hydrologic models, Journal of Hydrology, 204, 83-97, http://dx.doi.org/10.1016/S0022-1694(97)00107-8, 1998.

Zolezzi, G., Bellin, A., Bruno, M. C., Maiolini, B., and Siviglia, A.: Assessing hydrological alterations at multiple temporal scales: Adige River, Italy, Water Resources Research, 45, W12421, 10.1029/2008WR007266, 2009.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-121, 2017
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 6 March 2017

Hydrology and Earth System Sciences
Discussions

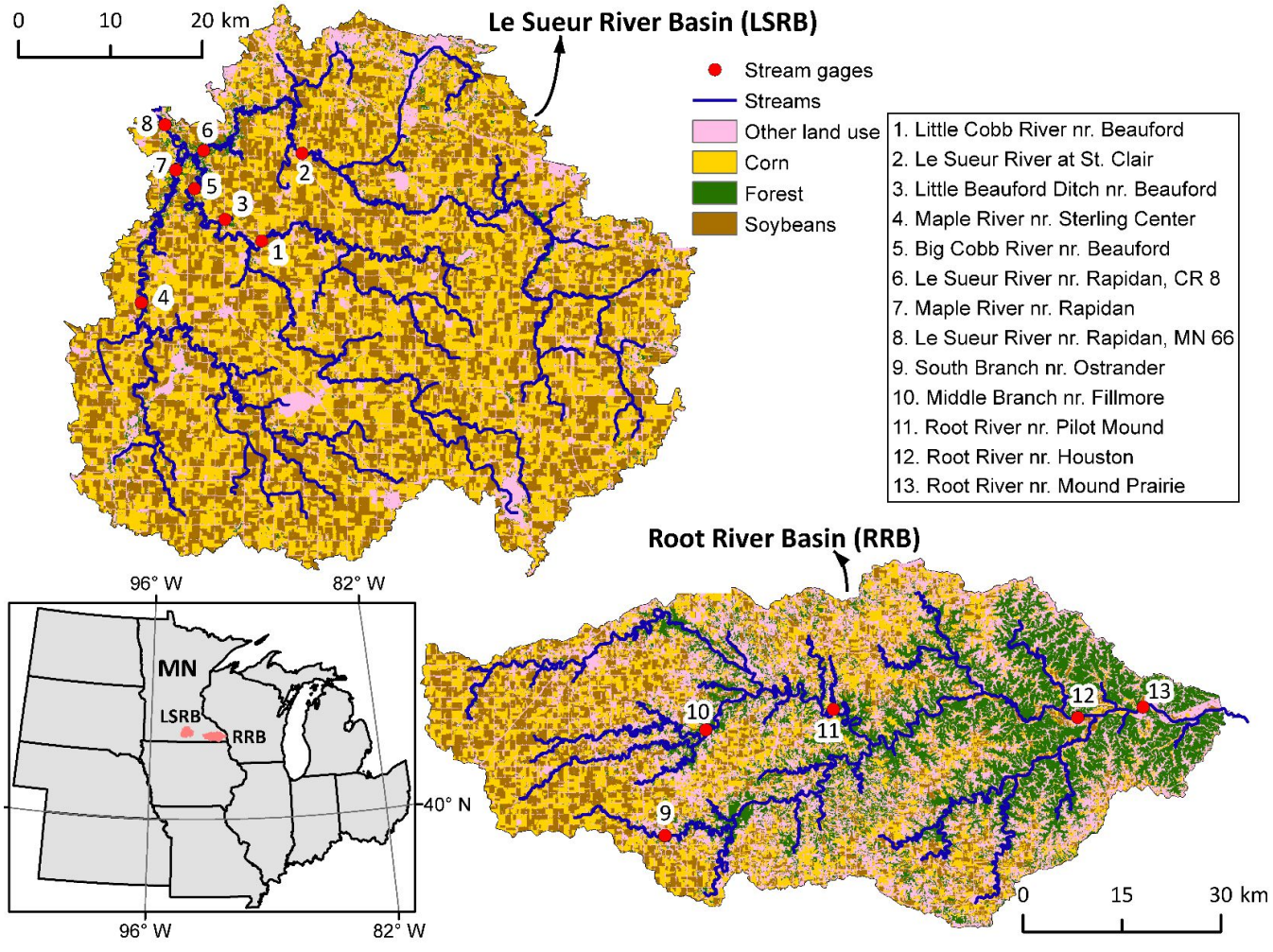

Figure 1: Map of midwestern US showing the two watersheds included in this study. Land use map of Le Sueur River Basin (LSRB) and Root River Basin (RRB) are overlaid with streamflow gages where the SWAT models was calibrated and validated. Land use from the year 2006 is shown for the two basins. Corn, soybean, forest (deciduous, evergreen and mixed forests) and all else (e.g., grass/ pasture, develop/ open space and others) are combined as other land use.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-121, 2017
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 6 March 2017

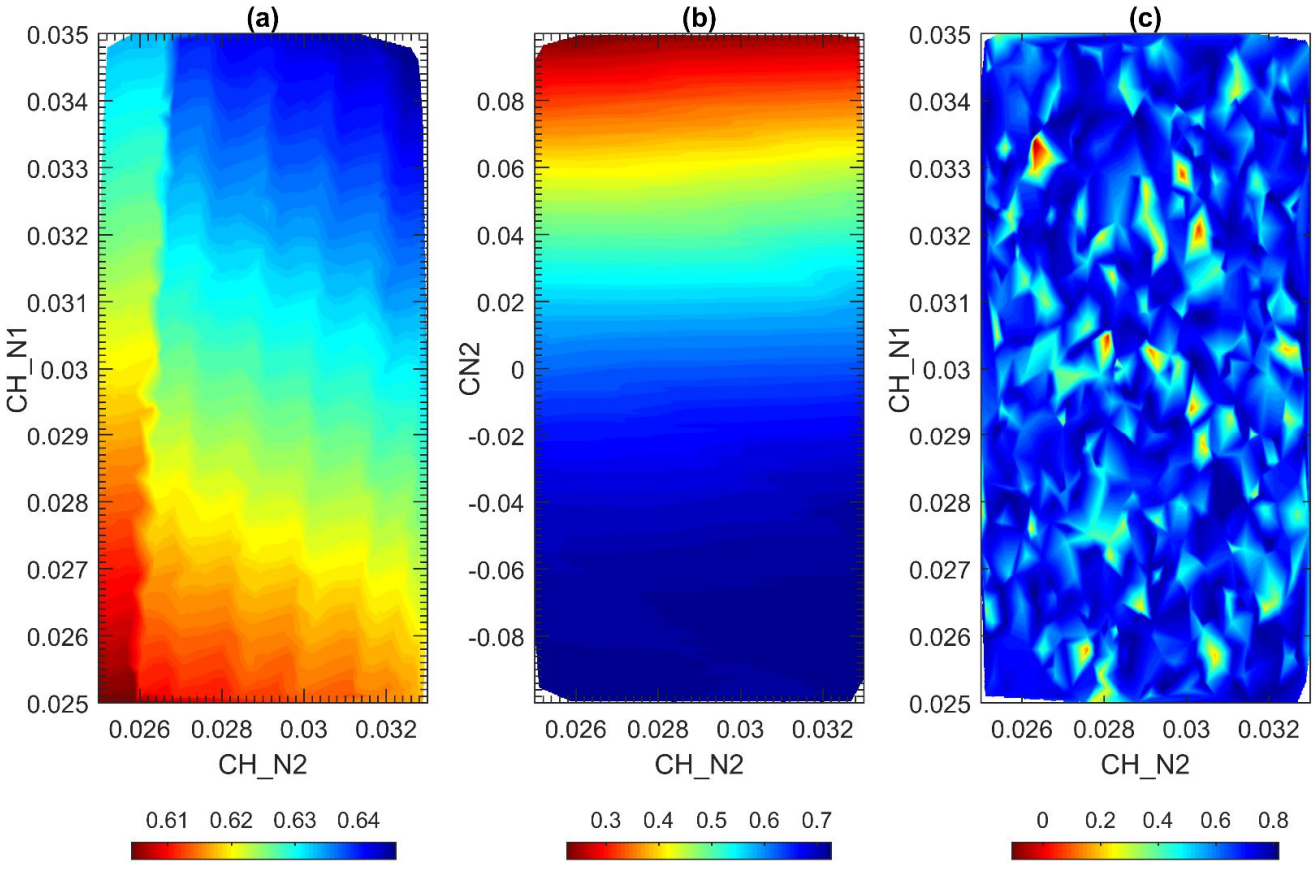

Figure 2: (a), (b) and (c) were generated from 1000 runs each using the SWAT model for the South Branch of the Root River (sub-model of the Root River SWAT model with same parametrization as the larger RRB model. (a) Contour plot of NSE (coded as color) value for Manning's "n" value for the tributary channels (CH_N1) and Manning's "n" value for the main channel (CH_N2). (b) Contour plot of NSE value for CN2 and Manning's "n" value for the main channel (CH_N2). (a) Contour plot of NSE value for Manning's "n" value for the tributary channels (CH_N1) and Manning's "n" value for the main channel (CH_N2) when 18 parameters were varied.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-121, 2017
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 6 March 2017

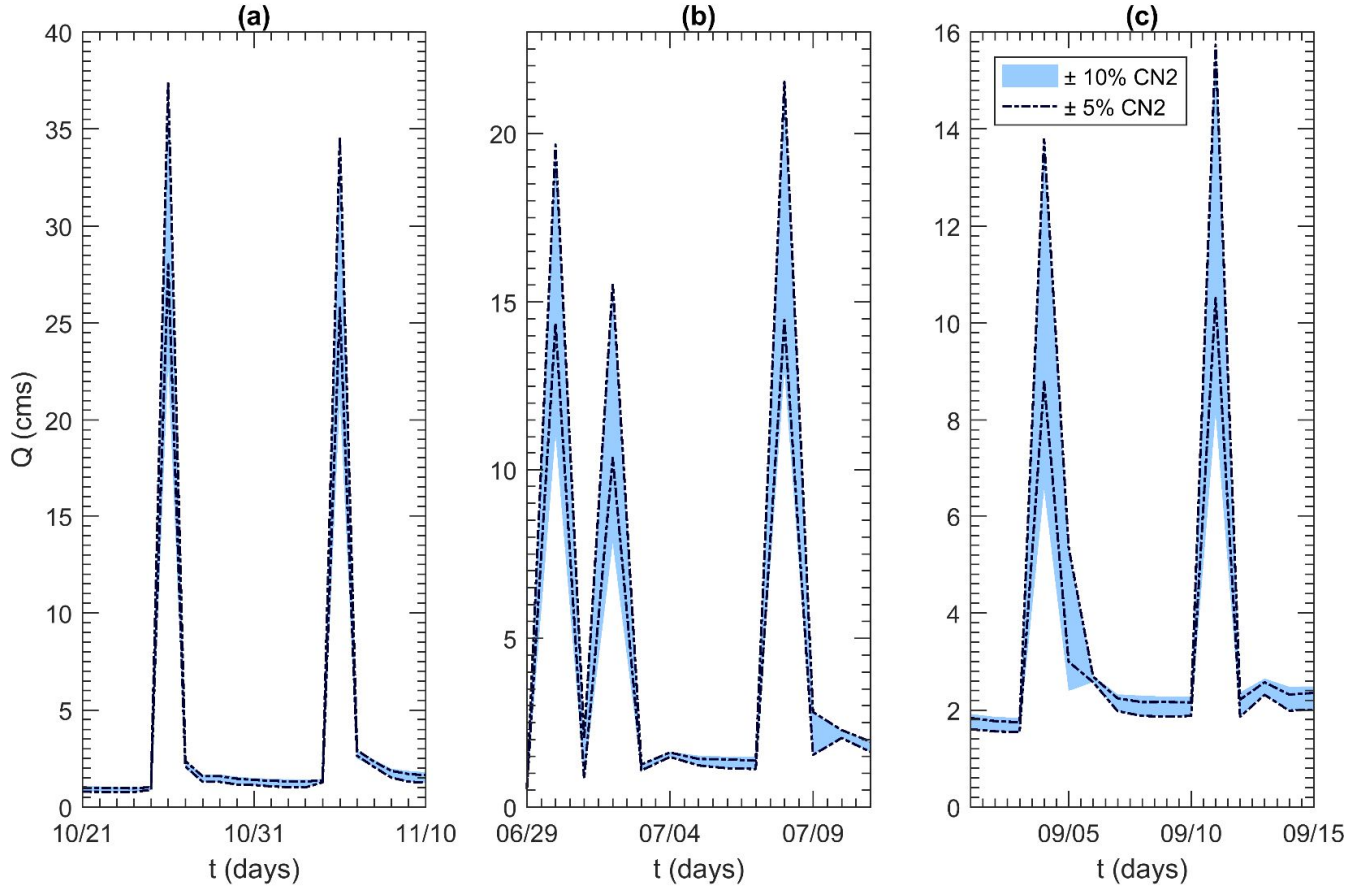

Figure 3: High sensitivity of CN2 is illustrated using 1000 runs from the South Branch of the Root River where CN was varied ±10% to mimic calibration and ±5% runs (1000 runs) to account for any potential model application to simulate a management practice. Only runs with NSE >0.6 are shown for both cases.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-121, 2017
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 6 March 2017

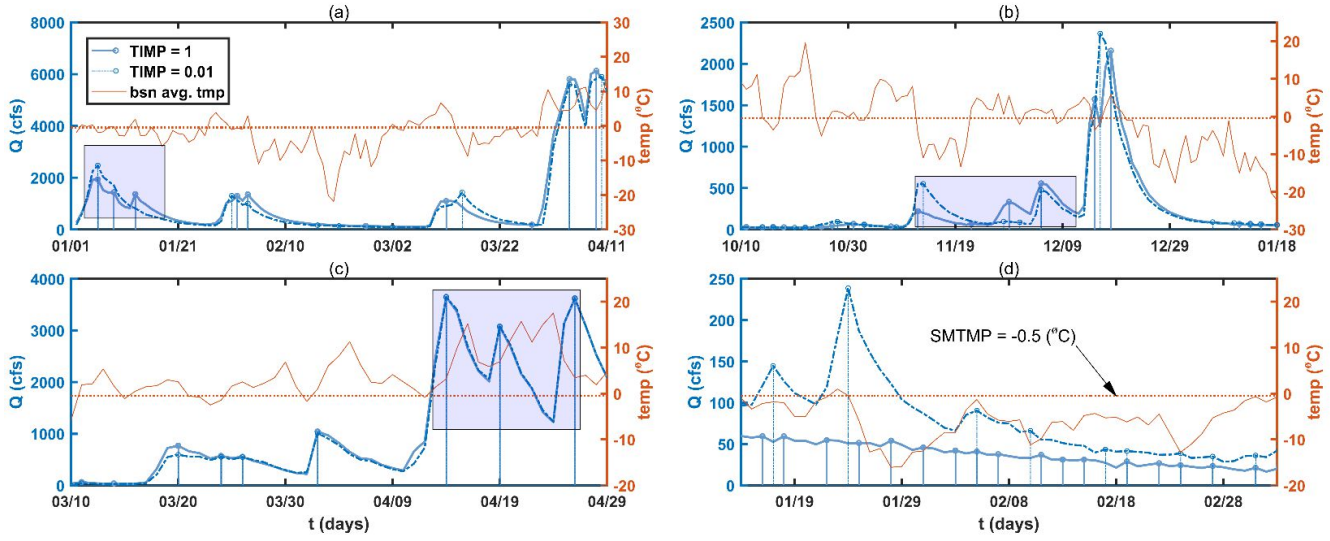

Figure 4: Daily average streamflow change from pre (TIMP = 1) to post (TIMP = 0.01) adjustment contrasted with basin average daily air temperature (derived from averaging daily air temperature from 175 subbasins). Dots indicate local peak flow values identified by the peak flow selection tool available in HydroME Toolbox. Dotted line represents the threshold temperature for snowmelt.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-121, 2017
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 6 March 2017

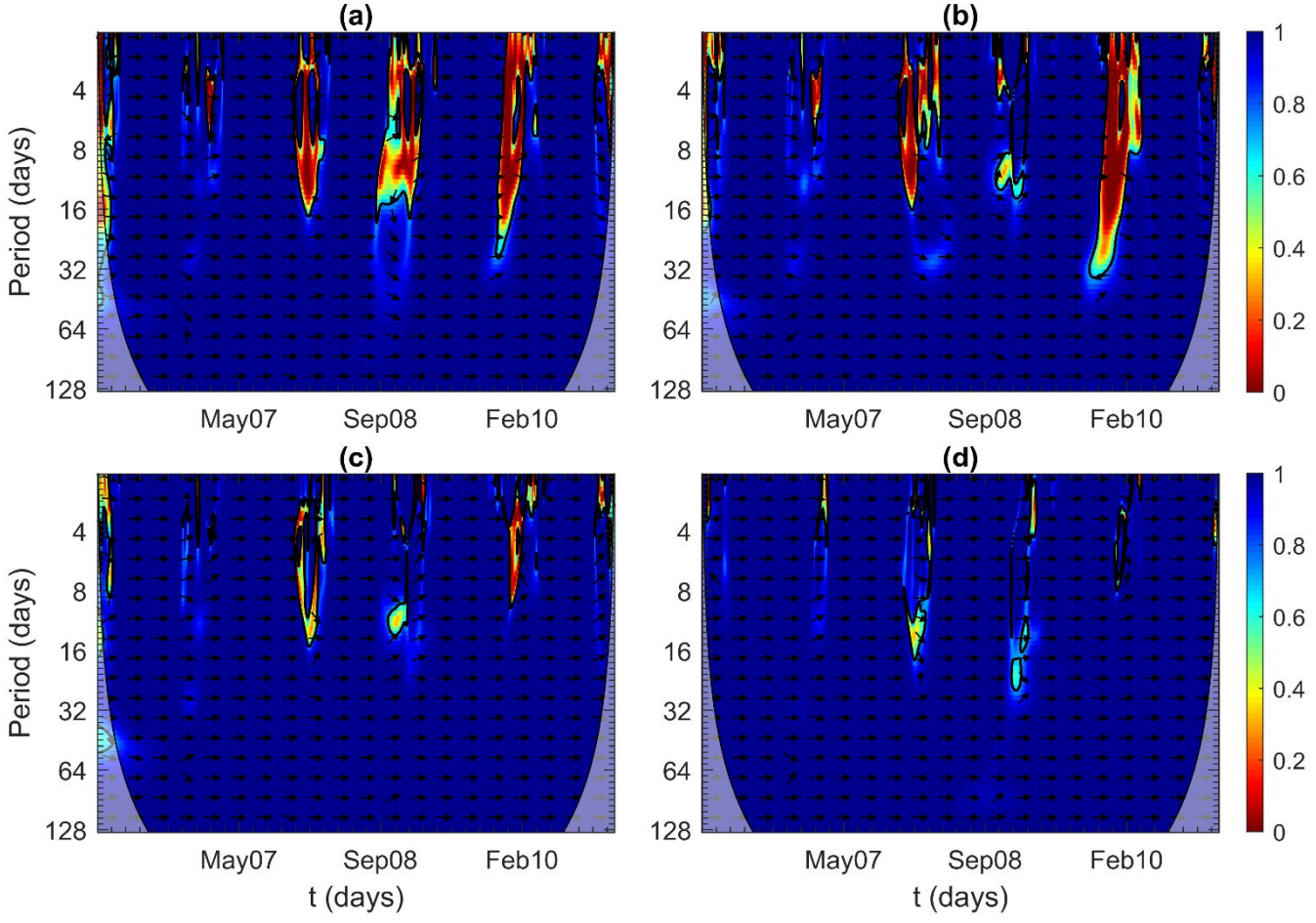

Figure 5: Dissimilarity in the time-scale domain between pre and post parameter adjusted streamflow signals are shown with the Wavelet Transform Coherence (WTC). The cone of influence is represented using a lighter shade (region where edge effects become dominant). The abscissa is time and the ordinate is the wavelet scale equivalent to the Fourier period. The coherence value is encoded by color. The thick black contour lines represent the 5% significance level against red noise. WTC for four commonly adjusted SWAT parameters to achieve calibration are shown, where (a) is TIMP, (b) is SFTMP, (c) is SMFMX and SMFMN and (d) is CN_FROZ. Change of streamflow shown in the time-scale domain resulting from one-at-a-time parameter adjustment. All four parameters are considered as pure calibration parameters in this paper.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-121, 2017
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 6 March 2017

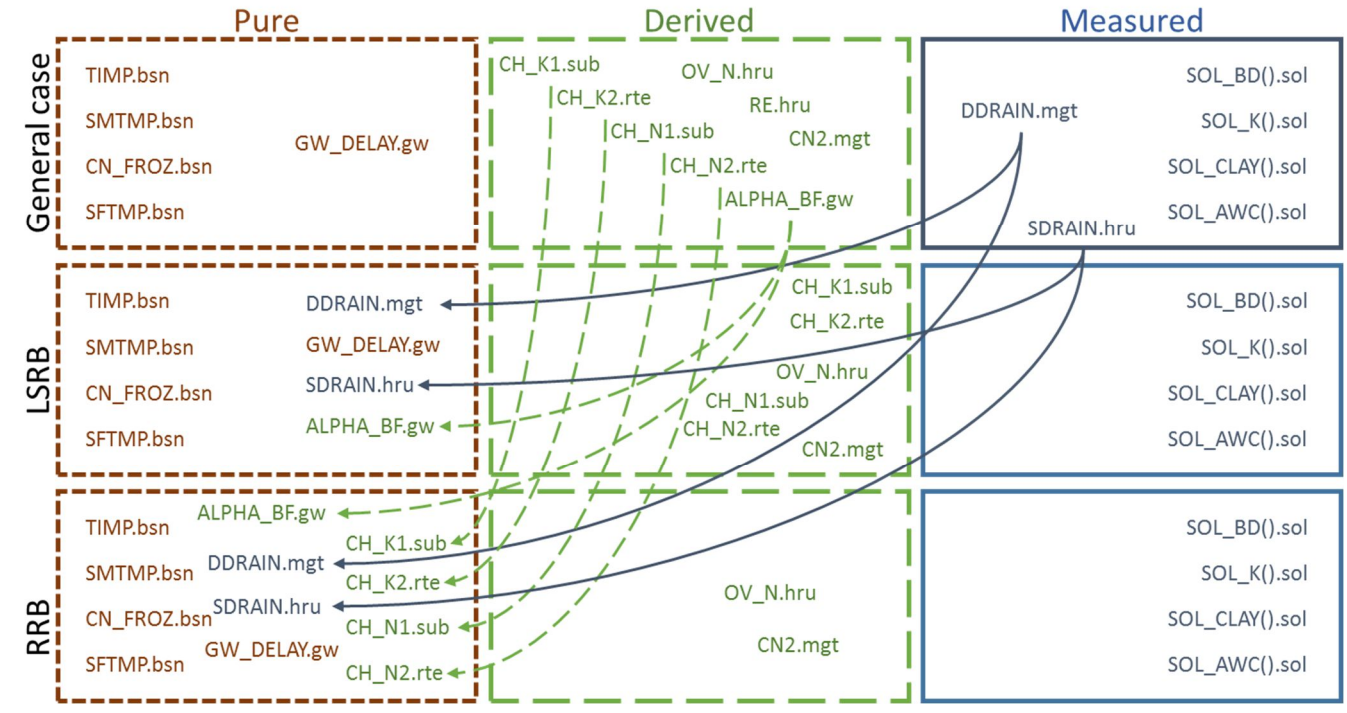

Figure 6: Flexible nature of parameter classification.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-121, 2017
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 6 March 2017

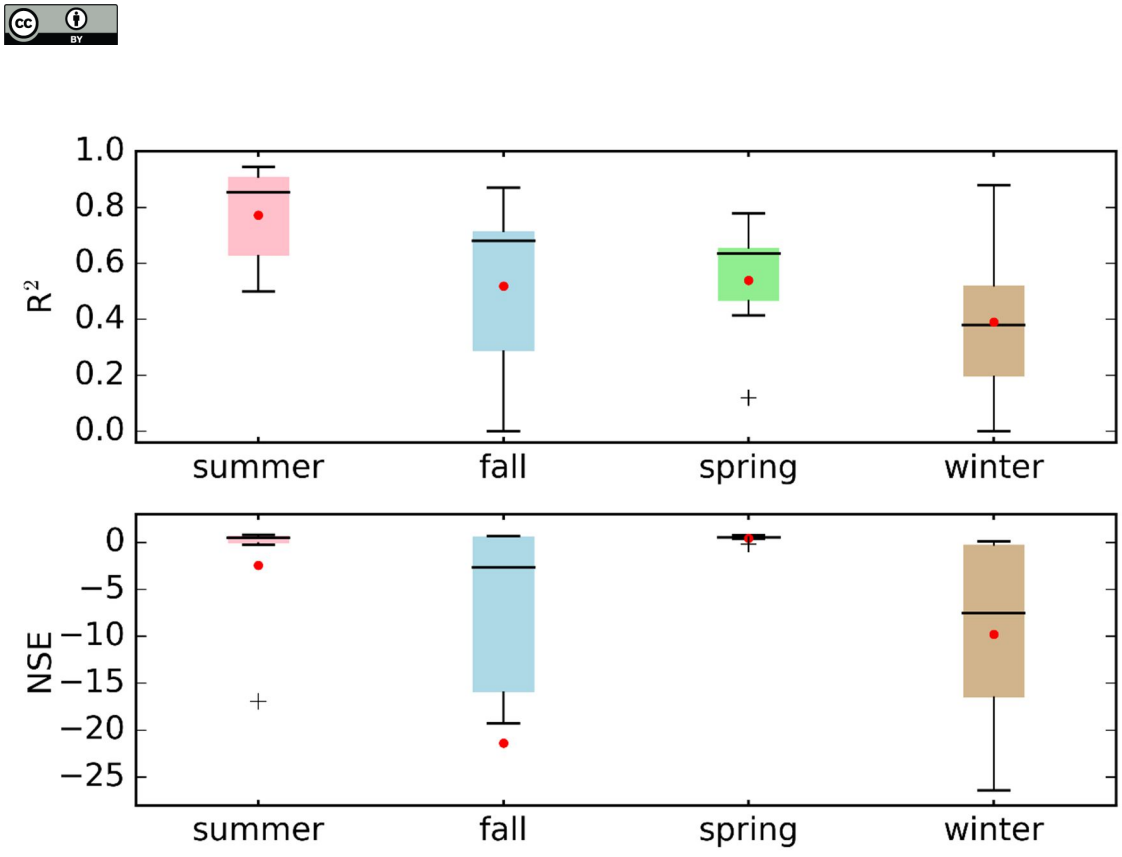

Figure 7: Seasonal NSE and R2 for the Root River Basin for the calibration period. Average overall NSE and R2 for the calibration period are 0.69 and 0.70, respectively. The box and whisker plot shows the eight number summary of the performance metrics. The height of the box portion is given by the interquartile range (IQR, 25th to 75th percentile) of the performance metrics NSE and R2. The horizontal bar is the median, red dot is the mean and the whiskers represent the 10th and 90th percentile values. The extreme values are shown with a "+".

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-121, 2017
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 6 March 2017

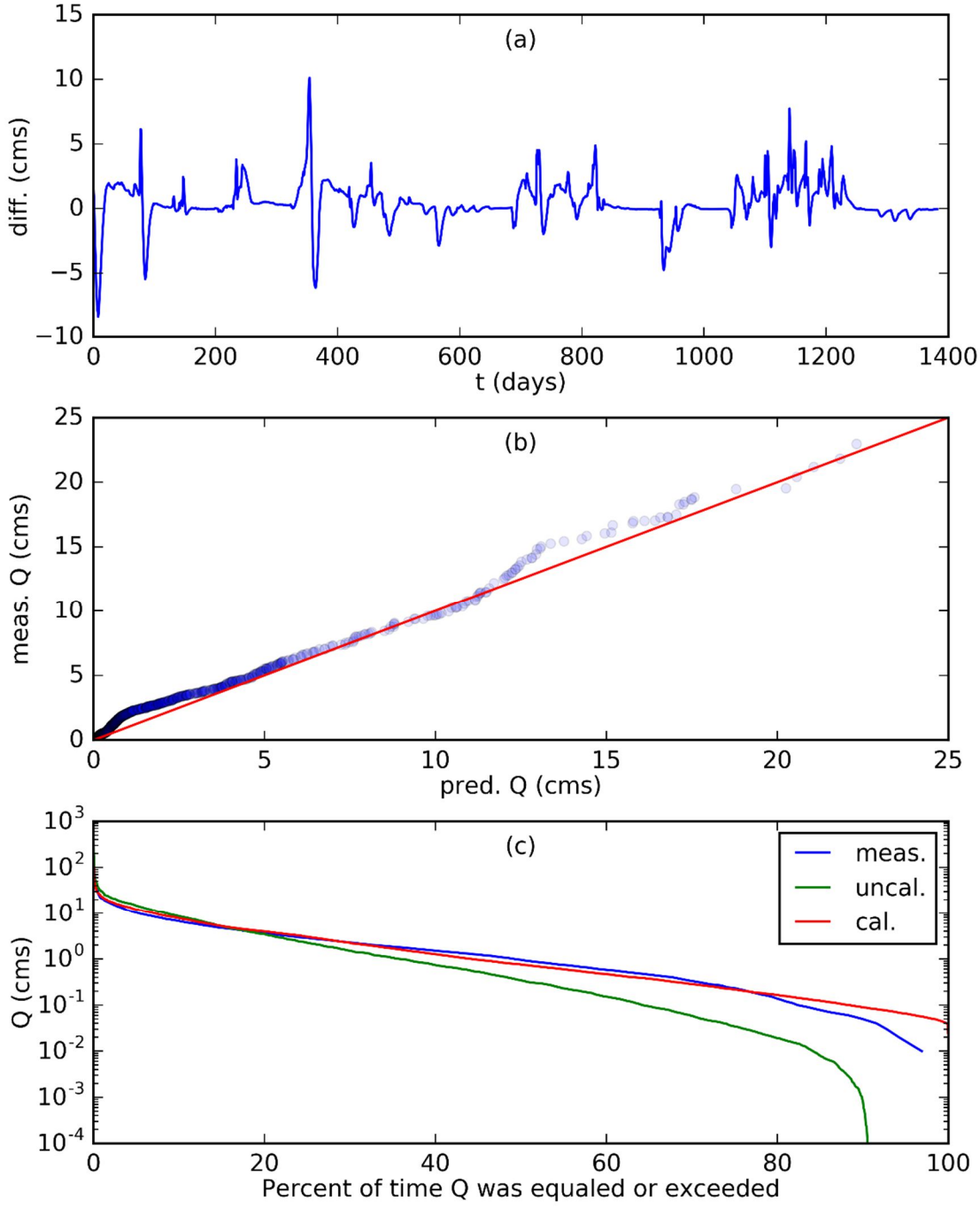

Figure 8: (a) Euclidian distance between the measured and predicted flow (b) Q-Q plot between the measured and predicted flows for the Little Cobb River in the Le Sueur River Basin. NSE = 0.78 for the data shown in the plot. The solid red line is the 1:1 reference line. (c) Flow Duration Curve for uncalibrated and calibrated model flows contrasted with measured flow for the Little Cobb River in the Le Sueur River Basin.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-121, 2017
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 6 March 2017

Hydrology and Earth System Sciences
Open Access
Discussions

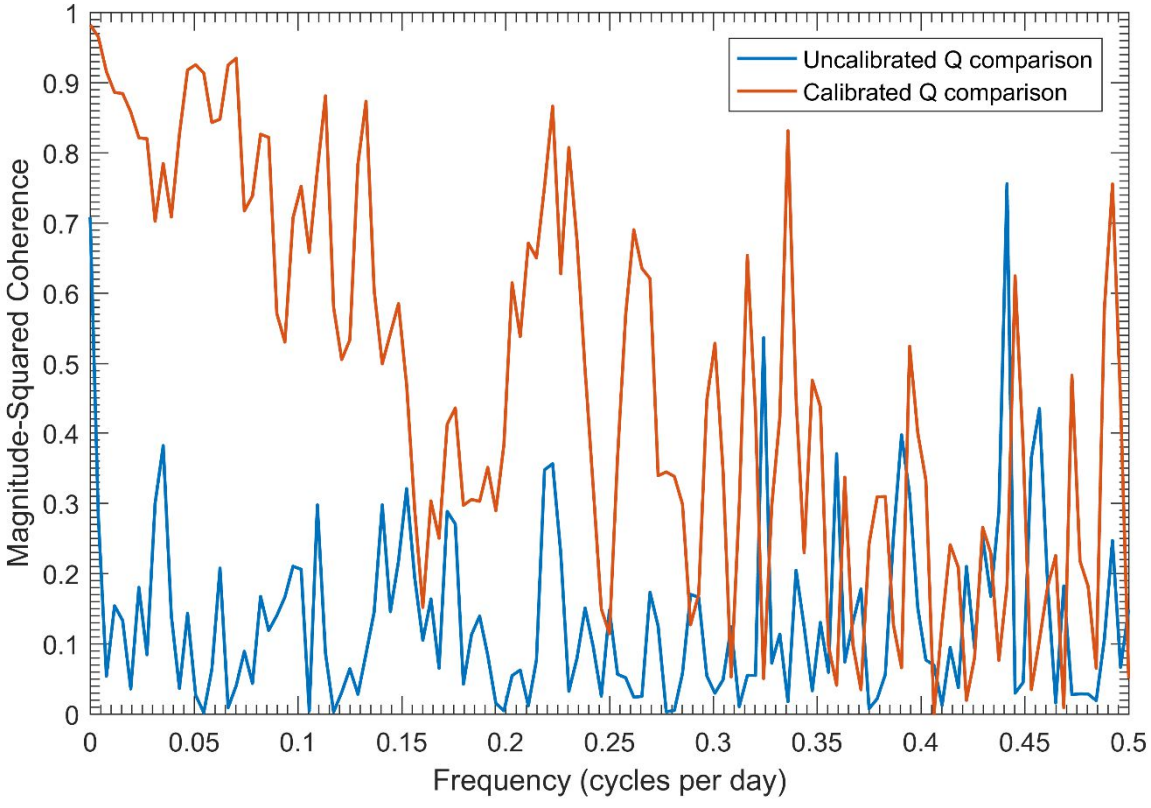

Figure 9: Magnitude squared coherence via Welch comparing measured streamflow with uncalibrated and calibrated streamflow. A value of 1 indicates the frequency is present in both the measured and predicted signals and a zero means it was present in one and not the other.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-121, 2017
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 6 March 2017

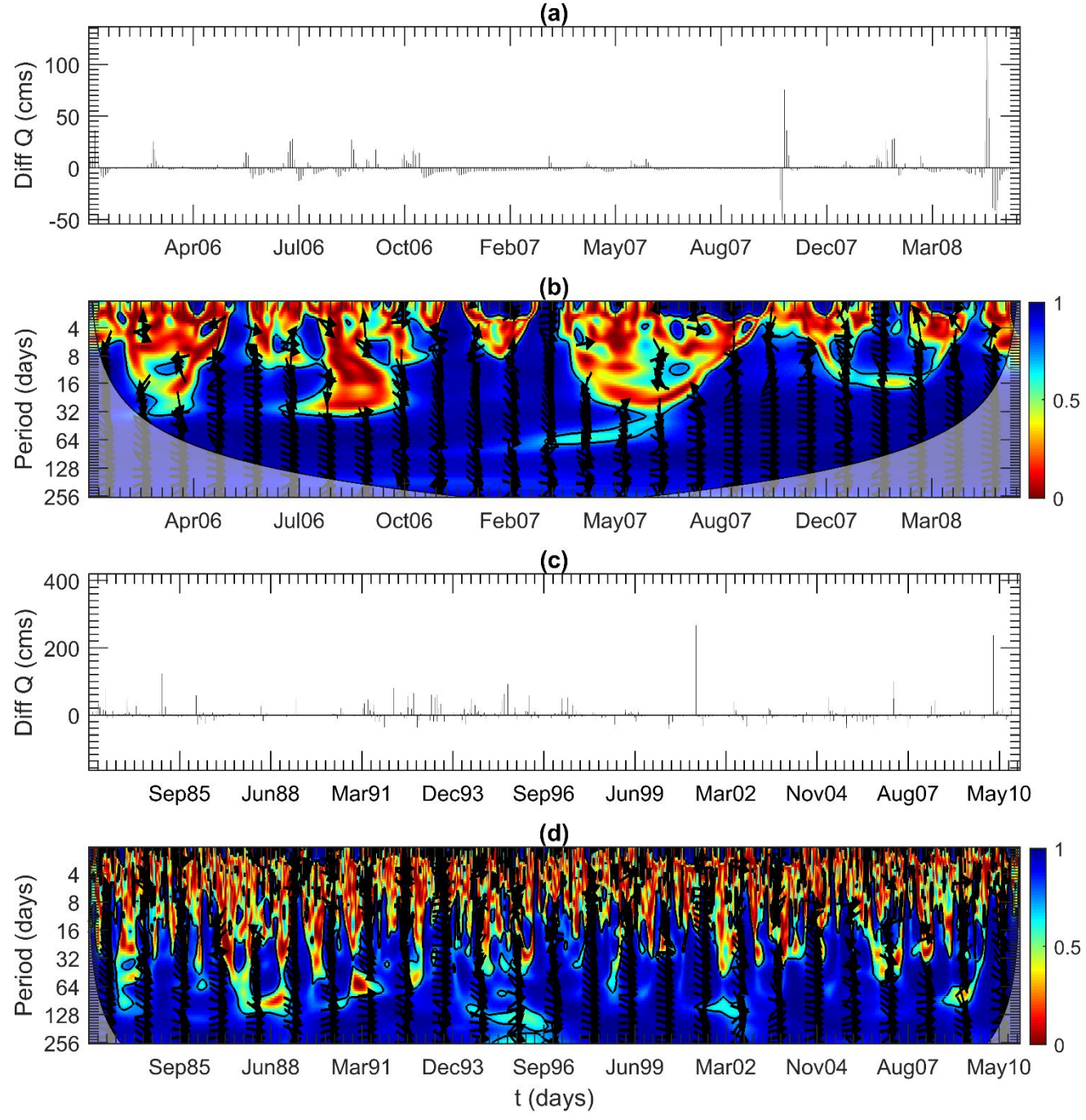

Figure 10: (a) and (c) Euclidian metric (between measured vs. simulated) for the Le Sueur River at St. Clair gage and Le Sueur River near Rapidan, MN 66 gage in the Le Sueur River Basin respectively. (b) and (d) Dissimilarity in time-scale domain between the measured and predicted signals are illustrated with the Wavelet Transform Coherence (WTC) for the Le Sueur River at St. Clair gage and Le Sueur River near Rapidan, MN 66 gage in the Le Sueur River Basin respectively. The cone of influence is represented using a lighter shade (region where edge effects become dominant). The abscissa is time and the ordinate is the wavelet

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-121, 2017
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 6 March 2017

Hydrology and Earth System Sciences
Open Access
EGU
Discussions

scale equivalent to the Fourier period. The coherence value is encoded by color. The low coherence in the ~<16 days period band is attributable to the inability of the model to capture higher frequency peak flows.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-121, 2017
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 6 March 2017

Hydrology and Earth System Sciences
Open Access
Discussions

Table 1: Lumped metrics (NSE and R^2) shown as commonly reported in the literature based on separation of data into calibration and validation data sets for RRB and LSRB at daily time step.

	Gage	Calibration	Validation	Calibration	Validation
		NSE		R^2	
Root River Basin	Middle Branch	0.60	0.56	0.70	0.73
	Houston	0.54	0.59	0.57	0.61
	Mound Prairie	0.68	0.52	0.72	0.61
	Pilot Mound	0.69	0.64	0.70	0.67
	South Branch	0.50	0.56	0.67	0.63
Le Sueur River Basin	Le Sueur River at St. Clair, CSAH28	0.67	0.72	0.67	0.74
	Le Sueur River at Rapidan, CR8	0.57	0.75	0.57	0.76
	LSR near Rapidan, MN66	0.59	0.73	0.60	0.74
	Little Beauford Ditch	0.54	0.50	0.55	0.62
	Little Cobb River near Beauford	0.74	0.79	0.76	0.79
	Big Cobb River near Beauford, CR16	0.83	0.84	0.84	0.85
	Maple River near Rapidan, CR35	0.72	0.69	0.73	0.70
	Maple River near Sterling Center, CR18	0.72	0.73	0.70	0.75