Advances in Water Resources 130 (2019) 129–146

Contents lists available at ScienceDirect

Advances in Water Resources

journal homepage: www.elsevier.com/locate/advwatres

Multivariate calibration of large scale hydrologic models: The necessity and value of a Pareto optimal approach

Akash Koppa*, Mekonnen Gebremichael, William W-G Yeh

Department of Civil and Environmental Engineering, University of California, Los Angeles, CA 90095, USA

ARTICLE INFO

Keywords:
Multivariate calibration
Bayesian calibration
Large scale hydrologic modelling
Pareto optimality

ABSTRACT

Multivariate calibration using measurements of multiple water balance components has emerged as a potential solution for improving the performance and realism of large scale hydrologic models. In this study we develop a novel multivariate calibration framework to rigorously test whether incorporation of multiple water balance components into calibration can result in sufficiently accurate (behavioral) solutions for all model responses. Unlike previous studies, we use Bayesian calibration to formally define limits of acceptability or error thresholds in order to distinguish behavioral solutions for each of the incorporated fluxes. We apply the framework in the Mississippi river basin for the calibration of a large scale distributed hydrologic model (Noah-MP) with different combinations of model responses - evapotranspiration (ET), soil moisture (SM), and streamflow (SF). The results of the study show that incorporation of additional fluxes and soil moisture (a storage variable) is not always valuable due to significant trade-offs in accuracy among the model responses. In our experiments, only ET and SF could be simulated simultaneously to a reasonable degree of accuracy. In addition, we quantify the trade-offs in accuracy between the model responses using the concept of Pareto optimality. We find that combining ET with other fluxes entails higher trade-offs in accuracy compared to either SM or SF. Unlike deterministic calibration, with the developed framework we are able to identify deficiencies in model parameterization that lead to significant trade-offs in accuracy, especially between ET and SM. We find that the parameters which are insensitive to individual model responses can influence the trade-off relationship between them.

1. Introduction

The widespread use of large scale hydrologic and land surface models (LSMs) in snow (Christensen and Lettenmaier, 2007; Li et al., 2017), drought (Leng et al., 2015; Sheffield et al., 2004), and climate change (Middelkoop et al., 2001; Cuo et al., 2013) studies necessitates critical examination of the adopted calibration methodologies. The general approach of calibrating the models with measurements of a single flux, typically streamflow, is considered inadequate for such studies that require other water balance components to be simulated accurately. Rakovec et al. (2016a) evaluate the performance of the mesoscale hydrologic model (mHM) calibrated with streamflow (SF) against observed evapotranspiration (ET), soil moisture (SM), and total water storage (TWS). The study concludes that calibrating hydrologic models with only streamflow may not be sufficient for accurate simulation of other water balance components. Wanders et al. (2014) calibrate the LISFLOOD hydrologic model with remotely sensed soil moisture datasets. The results of the study show that calibrating the hydrologic model with only soil moisture negatively affects the accuracy of the corresponding streamflow simulation (compared to streamflow-calibrated model results). López López et al. (2017) confirm the findings of the other studies; calibrating a hydrologic model with only ET or SM adversely affects the accuracy of streamflow simulation compared to streamflow-calibrated model results. In Zink et al. (2018), a land surface model calibrated with land surface temperature leads to higher errors in streamflow simulations compared to a streamflow-calibrated model.

Incorporating multiple fluxes and other water balance components into the calibration process has emerged as a consensus solution to address the adverse effects of single objective calibration. A number of different methods have been employed to calibrate hydrologic and land surface models with multiple fluxes, including stepwise calibration (López López et al., 2017; Sutanudjaja et al., 2013), ensemble Kalman filter (Wanders et al., 2014), and simultaneous calibration by combining objective functions (Rientjes et al., 2013; Rakovec et al., 2016b; Zink et al., 2018). Irrespective of the calibration strategy adopted, all the studies report improvements in the simulation of the added flux or storage component while maintaining the accuracy of the primary variable of interest. The improvements are also consistent across different water balance components incorporated into calibration, including

[*] Corresponding author.
E-mail address: akashkoppa@ucla.edu (A. Koppa).

https://doi.org/10.1016/j.advwatres.2019.06.005
Received 21 March 2019; Received in revised form 26 May 2019; Accepted 8 June 2019
Available online 9 June 2019

evapotranspiration (Rientjes et al., 2013; López López et al., 2017; Zink et al., 2018), soil moisture (Sutanudjaja et al., 2013; Wanders et al., 2014), and total water storage (Rakovec et al., 2016b). Although the enumerated studies provide evidence in favor of multivariate calibration, we identify shortcomings in these approaches that hinder comprehensive quantification of the value of incorporating additional fluxes.

First, multivariate calibration studies do not define any limits of acceptability or error thresholds to determine whether the model can simultaneously reproduce the incorporated water balance components, such as ET, SM, SF, and TWS, to a sufficient degree of accuracy. To illustrate the importance of defining limits of acceptability, consider the results of Rakovec et al. (2016b) wherein the addition of TWS estimates into calibration along with streamflow reduce the root mean square error (RMSE) of TWS simulations at negligible cost to the accuracy of streamflow simulation. A closer analysis of the results reveals that despite reduction in the RMSE of TWS, the absolute value of RMSE is still significantly large (Fig. 3. in Rakovec et al. (2016b)); whereas the RMSE of standardized anomalies of streamflow has a median of approximately 0.5, the RMSE of TWS is approximately 0.8 (reduced from 0.9 for the SF-calibrated model). Without defining a threshold for acceptable error, it is difficult to assess whether the reported reduction in TWS error is sufficient evidence to conclude that the incorporation of an additional flux actually improves the realism of the model.

Second, most studies consider the relationship among different water balance components to be complementary but all the results, with the exception of Wanders et al. (2014), point towards a trade-off relationship. By definition, a complementary relationship would mean that incorporation of additional fluxes and other water balance components improves the accuracy of all the incorporated fluxes. The objective functions of calibration and the calibration methodologies (such as stepwise calibration) are constructed to reflect the assumption of a complementary relationship. Even in Wanders et al. (2014) the improvement in SF accuracy when SM is incorporated, compared with a streamflow-calibrated model, is limited to small catchments. In addition, the results of multivariate calibration rarely are compared with results of models calibrated only with the additional flux. Such a comparison would help in quantifying the potential trade-offs in simulating the two water balance components accurately. For example, in Rakovec et al. (2016b) the model is not calibrated with only TWS, which would help understand the trade-off in TWS accuracy required to achieve acceptable SF accuracy. In studies where all the calibration cases are reported, there are significant trade-offs among the different fluxes or state variables considered for calibration (Rientjes et al., 2013). Even in Zink et al. (2018), where the objective function is designed to produce a compromise solution between the SF and ET fluxes, there is no discussion of either the magnitude of trade-off in the accuracy of ET or of whether such trade-offs are within acceptable limits.

Third, the trade-off relationship among the simulated water balance components implicit in the results of multivariate calibration studies may be a consequence of deficiencies in model structure and parameterizations (Fenicia et al., 2007; Hogue et al., 2006). However, in the calibration strategies adopted in most studies, including the assumption of a complementary relationship among the fluxes and/or other water balance components, combining objectives and lack of a definition of error thresholds prevent any meaningful diagnoses of the model. For example, the limitations of the stepwise calibration methodology for identifying deficiencies in model structure and parameterizations are well known (Fenicia et al., 2007). Additionally, most multivariate calibration studies are deterministic and hence are inappropriate for studying the differences in optimal parameter sets between univariate and multivariate calibration cases, as they do not address the issue of equifinality (Beven, 1996, 2001). Even in studies that use stochastic methods such as ensemble Kalman filter (Wanders et al., 2014), there is little discussion on how parameters behave between different calibration cases.

In this study, we combine a formal Bayesian calibration approach with the concept of Pareto optimality to address the issues detailed above. We utilize a formal Bayesian calibration approach to define the limits of acceptability or error thresholds in order to distinguish between behavioral and non-behavioral solutions (Beven, 2006; Vrugt et al., 2009a) for each of the incorporated water balance components. 'Behavioral' solutions are model parameter sets that result in errors that are within a defined threshold or limit with respect to a specific simulated response (for example ET or SM). If a trade-off relationship does exist among the incorporated water balance components, as opposed to a complementary relationship, the concept of Pareto optimality would help in understanding the extent to which the accuracy of a particular water balance component can be improved without affecting, to an unreasonable degree, the accuracy of other water balance components. In addition, Pareto optimal solutions are unbiased by any subjective weights given to any particular flux or water balance component over another, unlike simultaneous calibration strategies (Gupta et al., 1998). Hence, we use Pareto optimality-based calibration to create a set of non-dominated solutions that characterize the trade-offs among the incorporated variables. We develop a multivariate calibration framework that combines behavioral solutions from Bayesian calibration and multivariate calibration solutions to address the following research questions: 1) Does incorporation of multiple fluxes and other water balance components into calibration produce parameter distributions that are behavioral with respect to all water balance components considered for calibration? 2) For a given large scale hydrologic model, what is the extent of trade-off, if any, in accurate simulations of multiple water balance components considered for calibration? 3) Can behavioral and multivariate calibration solutions help identify deficiencies in hydrologic model parameterization that lead to trade-offs in the accurate simulations of multiple variables?

2. Methodology

2.1. Conceptual framework

Consider a hydrologic model,

$$O = \mu(\theta, I) \tag{1}$$

where O is a matrix consisting of model output or responses (such as evapotranspiration, soil moisture, streamflow, etc.); I is a matrix consisting of model input (meteorological forcings such as precipitation, air temperature, etc.); μ represents the mathematical structure of the hydrologic model, typically a deterministic or stochastic function, such that μ: $I \mapsto O$; and $\theta$is a vector of model parameters (Kavetski et al., 2006). Given a matrix of observations, $\widehat{O}$, a measure, L, can be defined as

$$L\big(E(\theta) = O(\theta) - \hat{O}\big) = S \tag{2}$$

where E is the error residual matrix, L is a measure or metric that preserves the information contained in the residuals (such as mean absolute error or root mean square error), and $S \in (-\infty, \infty)$ is some scalar quantity that represents the value of L.

Assuming that the model structure (μ) and the upper and lower limits of the parameters (θ) are fixed, the feasible bounds for Eq. (2), termed as the objective space, can be defined (Gupta et al., 1998) (A conceptual representation of a feasible objective space for two objectives (L_1 and L_2) is shown in Fig. 1.) In traditional model calibration, an optimal parameter set, θ^*, is identified by minimizing Eq. (2) ($L_1^* = L(\theta_1^*)$ and $L_2^* = L(\theta_2^*)$ are the optimal values for objective 1 and 2 in Fig. 1). Using formal or informal Bayesian approaches, it is also possible to identify sets of parameters, θ^b, that result in behavioral solutions, based on a defined cut-off threshold (Vrugt et al., 2009a). ($L_1^b = L(\theta_1^b)$ are behavioral solutions for objective 1, represented by the objective space to the left of the cut-off threshold e_1, and $L_2^b = L(\theta_2^b)$ are the behavioral solutions for objective 2, represented by the objective space below e_2 in Fig. 1). To quantify the trade-offs among multiple objectives (L_1 and L_2in this example), the concept of Pareto optimality can be used. This results in a set of parameters, θ^p, that give rise to non-dominated or Pareto optimal

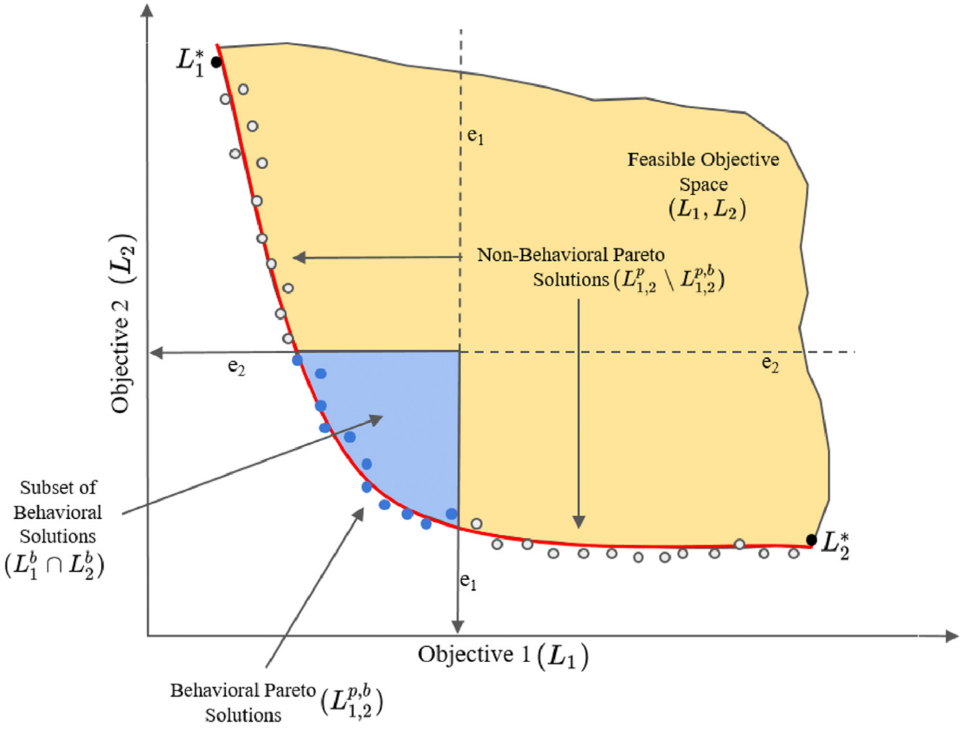
L_1^*
e1
Feasible Objective Space (L_1, L_2)
Non-Behavioral Pareto Solutions $(L_{1,2}^p \setminus L_{1,2}^{p,b})$
Objective 2 (L_2)
e2
e2
Subset of Behavioral Solutions $(L_1^b \cap L_2^b)$
e1
L_2^*
Objective 1 (L_1)
Behavioral Pareto Solutions $(L_{1,2}^{p,b})$

Fig. 1. Conceptual framework of the methodology adopted in this study (adapted from Efstratiadis and Koutsoyiannis (2010)).

solutions for the objectives considered ($L_{1,2}^p = L(\theta_{1,2}^p)$ are the Pareto optimal solutions for two objectives, represented by the points along the red line in Fig. 1).

In this study, the measure L is the root mean square error (RMSE) and the model responses, O, considered for calibration are evapotranspiration (ET), soil moisture (SM), and streamflow (SF). To address the first research question, we test the following hypothesis:

$$L_{1,2,\ldots n}^{p,b} = L_{1,2,\ldots n}^{p} \cap \left(L_1^b \cap L_2^b \ldots \cap L_n^b\right) \neq \emptyset \tag{3}$$

where $L_{1,2,\ldots n}^{p}$ is the non-dominated or Pareto optimal solutions for n objectives, $L_1^b \cap L_2^b \ldots \cap L_n^b$ is the intersection of behavioral solutions for n objectives (blue space in Fig. 1 for two objectives), $L_{1,2,\ldots n}^{p,b}$ are solutions that are both non-dominated and behavioral with respect to n objectives (blue points in Fig. 1) and ∅ represents an empty set. In other words, for incorporation of multiple water balance components in hydrologic model calibration to be considered valuable, it should be possible to identify a set of parameters that result in both Pareto optimal and behavioral solutions for all the water balance components considered in calibration. We note that in defining the hypothesis, we have considered only non-dominated or Pareto optimal solutions in the multi-objective space (blue points in Fig. 1). However, any solution within the behavioral limits, non-dominated or dominated, can be considered as valid multivariate calibration solution. In such cases, the hypothesis to be tested reduces to $(L_1^b \cap L_2^b \ldots \cap L_n^b) \neq \emptyset$. In other words, the intersection of behavioral solutions for n objectives in the multivariate space (blue space in Fig. 1) must be a non-empty set. In this study, we only focus on the Pareto optimal solutions to test the hypothesis (Eq. 3) and analyze the trade-off among accurate simulations of multiple water balance components. In addition, the behavioral solutions from Bayesian calibration and the Pareto optimal solutions are discrete (unlike the representation in Fig. 1). Therefore, the probability of finding an intersection between the two is extremely low. Hence, the non-dominated solutions, $L_{1,2,\ldots n}^{p}$, that lie within the error thresholds or limits of acceptability $\{e_1, e_2, \ldots e_n\}$ are considered as $L_{1,2,\ldots n}^{p,b}$ (Fig. 1 provides an example for two objectives). We test the above hypothesis (Eq. 3) for different pairs of model responses ($n = 2$): (1) ET and SM, (2) ET and SF, and (3) SM and SF.

In summary, the developed multivariate calibration methodology consists of the following steps: (1) Define the limits of acceptability or cutoff error thresholds (RMSE in this study) for the model responses incorporated into calibration (ET, SM, and SF in this study), (2) Determine the Pareto optimal solutions for the combination of model responses (ET-SM, ET-SF, SM-SF in this study), (3) Test the hypothesis (Eq. 3), and diagnose deficiencies in model parameterization using behavioral and the Pareto optimal solutions (Fig. 2).

2.2. Defining limits of acceptability for individual model responses

To define the limits of acceptability, e1 and e2, we adopt a formal Bayesian approach to derive the posterior distribution of model error (RMSE) and parameters for individual model responses (ET, SM, and SF). Specifically, we utilize the Differential Evolution Adaptive Metropolis (DREAM) Markov Chain Monte Carlo (MCMC) scheme (Vrugt et al., 2009b, 2008) that has been applied to Bayesian or uncertainty-based calibration of hydrologic (Shafii et al., 2014) and hydrogeologic (Laloy et al., 2013) models. There is a specific reason for using DREAM in this study. Unlike informal Bayesian approaches such as Generalized Likelihood Uncertainty Estimation (GLUE) (Beven and Binley, 1992), the definition of cutoff thresholds to distinguish behavioral solutions is not subjective. Instead of rejecting solutions based on a subjectively defined threshold, DREAM uses a formal likelihood function to assign probabilities to all the solutions that form the final converged posterior distribution. As a result, a specific quantile of the sampled probability distribution can be considered the cutoff threshold for distinguishing behavioral solutions (Vrugt et al., 2009a).

In this study, we assume no apriori knowledge about the value of error (RMSE) thresholds for any of the model responses (ET, SM, and SF). Instead, we determine a set of limits of acceptability corresponding to 10%, 25%, 50%, 75%, 90%, 95% and 99%, quantiles from the posterior distribution of RMSE, derived using DREAM for ET, SM, and SF. Note that the likelihood function in DREAM considers error residuals and not RMSE to determine the posterior distribution of parameters. Recent advances such as approximate Bayesian computation has enabled the use of summary statistics and error metrics (such as RMSE) for diagnostic model calibration (Vrugt and Sadegh, 2013; Sadegh and Vrugt, 2014)

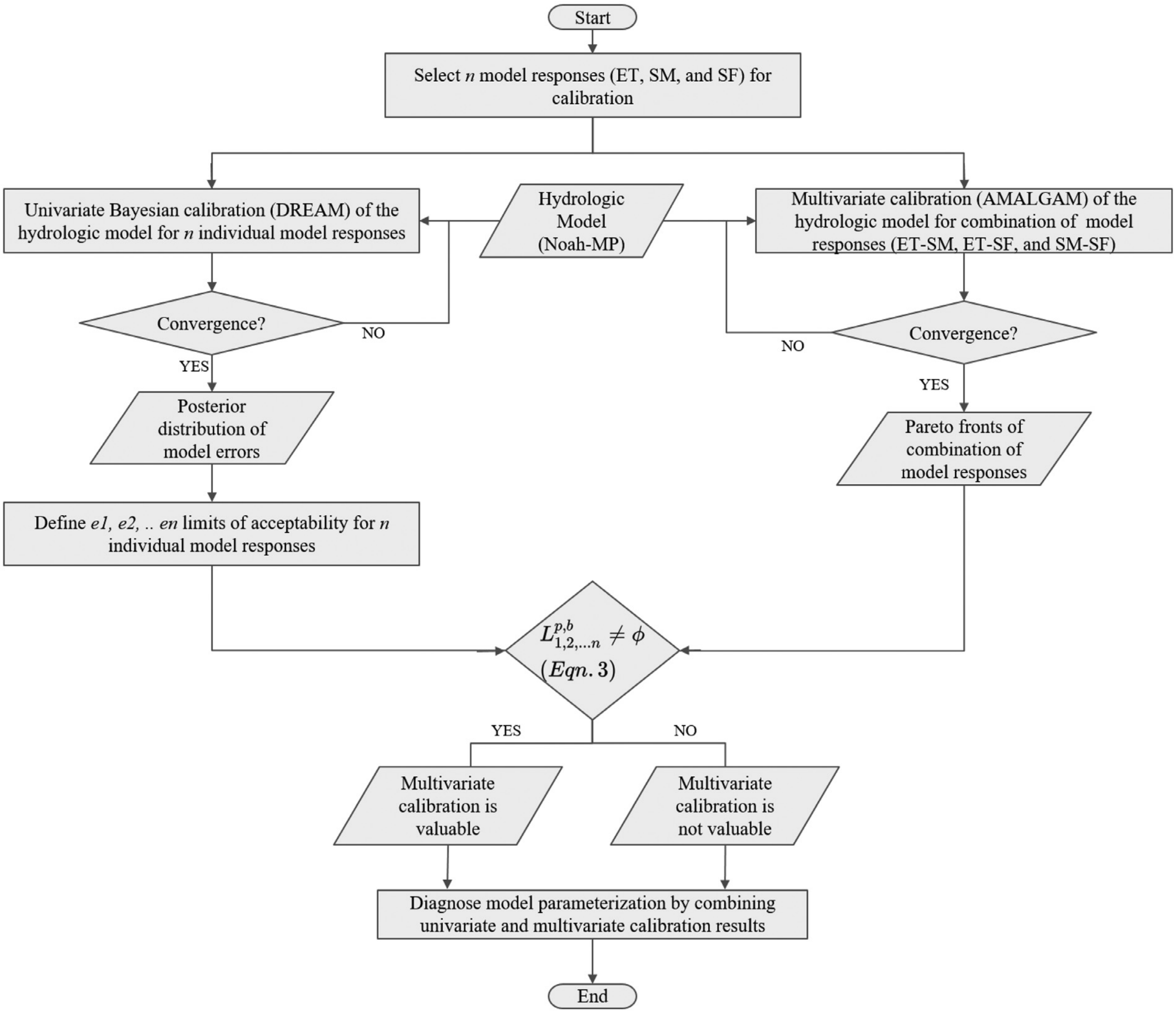
Start
Select n model responses (ET, SM, and SF) for calibration
Univariate Bayesian calibration (DREAM) of the hydrologic model for n individual model responses
Hydrologic Model (Noah-MP)
Multivariate calibration (AMALGAM) of the hydrologic model for combination of model responses (ET-SM, ET-SF, and SM-SF)
Convergence?
NO
YES
Posterior distribution of model errors
Define $e1, e2, .. en$ limits of acceptability for n individual model responses
Convergence?
NO
YES
Pareto fronts of combination of model responses
$L^{p,b}_{1,2,\ldots n} \neq \phi$
$(Eqn. 3)$
YES
NO
Multivariate calibration is valuable
Multivariate calibration is not valuable
Diagnose model parameterization by combining univariate and multivariate calibration results
End

Fig. 2. A schematic diagram of the multivariate calibration framework developed in this study.

and evaluation (Gupta et al., 2008). We do not use these methods in this study, as they require apriori definition of the limits of acceptability, similar to GLUE. We note that the DREAM solutions in this study are derived using residual and likelihood-based fitting methods, whereas the Pareto optimal solutions are derived using RMSE as the objective function (described below).

2.3. Pareto optimal solutions for combination of model responses

The value of using the concept of Pareto optimality for calibration of hydrologic models is well documented (Gupta et al., 1998). Studies have focused on identifying the best objective functions for improving streamflow calibration (Efstratiadis and Koutsoyiannis, 2010), developing systematic multi-objective calibration frameworks (Madsen, 2003) and analyzing the resulting Pareto fronts (Khu and Madsen, 2005). Multi-objective methods have been applied at small scales to constrain land surface model parameters (Gupta et al., 1999) and evaluate model performance and parameter behavior (Hogue et al., 2006). However, the utility of such an approach for multivariate calibration and diagnosis of large scale hydrologic models has received relatively less attention. In this study, we use A Multi-Algorithm Genetically Adaptive Multiobjective (AMALGAM) algorithm (Vrugt and Robinson, 2007) and RMSE as the objective function to derive non-dominated solutions for the following combinations of model responses: 1) ET and SM (ET-SM), 2) ET and SF (ET-SF), and 3) SM and SF (SM-SF).

2.4. Hypothesis testing, trade-off analysis and model diagnosis

To understand whether the models can accurately simulate multiple water balance components, we combine the behavioral solutions from DREAM and the Pareto optimal solutions from AMALGAM as detailed above. Specifically, we test the hypothesis defined in Eq. 3; the set of non-dominated solutions derived for different combinations of water balance components (ET-SM, ET-SF, and SM-SF) is a non-empty set for a particular behavioral limit (10%, 25%, 50%, 75%, 90%, 95% and 99% quantiles from the posterior distribution of RMSE). If a particular combination of water balance components has at least one Pareto optimal point within a stricter definition of error threshold compared to another combination, then the model is better at simulating the former combination of water balance components together. For example, consider that the ET-SM combination has at least one Pareto optimal solution within the 50% quantile error thresholds. On the other hand, consider that the ET-SF combination has at least one Pareto optimal solution within the 25% quantile error thresholds. In such a case, it can be concluded that it is valuable to incorporate ET and SF together compared to ET and SM.

To study the trade-off in the accuracy of the simulated model responses, we analyze the Pareto front qualitatively and quantitatively. Qualitatively, a well-defined Pareto front implies that the incorporated model water balance components exhibit a trade-off relationship as opposed to a complementary relationship. Quantitatively, the range of the objectives and the slope of the Pareto-front can help compare different

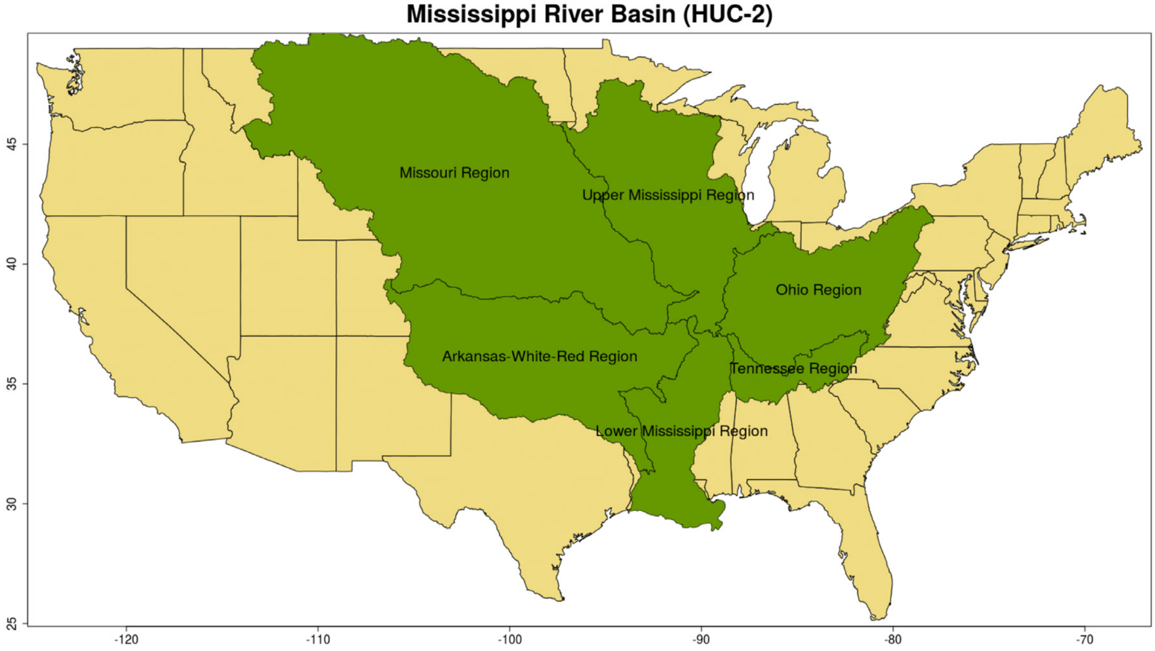

Fig. 3. A map of the Mississippi basin showing the six USGS HUC-2 basins.

Pareto fronts in order to understand the trade-off relationship between the different combinations of water balance components. Specifically, we define ‘magnitude of trade-off’ as the increase in the error of a specific model response required to affect a unit decrease in the error of the additional water balance component. We calculate and compare the average, maximum, and minimum magnitude of trade-off for each of the three multivariate calibration cases from the slopes of the Pareto front.

To diagnose the reasons for trade-offs in accuracy among different model responses, we study the differences in model parameter distributions among 1) behavioral solutions of individual responses, 2) behavioral solutions of individual model responses and behavioral Pareto optimal solutions, and 3) behavioral solutions of individual responses and Pareto optimal solutions that are not behavioral for any model response. We compare the empirical cumulative distribution functions (ECDFs) and quantiles of parameters to help identify parameters that most affect model behavior when additional water balance components are incorporated. We quantify the difference between the PDFs of the parameters using Hellinger’s distance, H, a statistical distance measure defined as

$$H(P,\ Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{\{i=1\}}^{\{k\}}\left(\sqrt{p_1}-\sqrt{q_i}\right)^2} \tag{4}$$

where, $P = (p_1,p_2,\ldots p_k)$ and $Q = (q_1,q_2,\ldots q_k)$ are discrete probability distributions. We determine the Kolmogorov-Smirnov (KS) test statistic to compare the ECDFs of different parameter distributions. The KS test statistic determines the maximum distance between two ECDFs. In addition, we examine how well the objective functions in the Pareto optimal solutions are correlated with the corresponding parameter sets. This will help us map the changes in parameter values along the Pareto front and identify parameters that influence the trade-off relationship between the objectives.

3. Experiment design

3.1. Study area and time period

To simulate large scale hydrologic studies, we choose the Mississippi basin in the United States as the study region. The basin covers an area of about 3.3 million sq. km. and six USGS HUC-2 (Hydrologic Unit Code) basins (Fig. 3). The average temperature over the basin is about 12 °C and the annual average rainfall is estimated as 800 mm (Cai et al., 2014). Cai et al. (2014) also classify the Ohio and Tennessee regions as wet regions, the Missouri basin as dry and the Upper Mississippi region as a transitional region between wet and dry. We use historical data from the year 2004 for calibration and data from the year 2005 for validation. Employing a monthly time step, 72 streamflow data points and approximately 63,000 ET and SM data points are used for calibration. The reasons for selecting a single year for calibration and validation of the model are two-fold: 1) The study is a calibration experiment that does not seek to produce the best-performing hydrologic model for the Mississippi river basin. Rather, the primary aim is to rigorously test whether large scale hydrologic models can behaviorally simulate multiple water balance components and study the reasons behind the trade-offs, if any, among accurate simulations of multiple water balance components. 2) The forward hydrologic model used in the study is computationally expensive.

3.2. Observational data

To simulate studies that use a sparse network of streamflow gauges for the calibration of hydrologic models, we use the computed monthly runoff for the six HUC-2 basins sourced from USGS. For calibrating the hydrologic model with remotely sensed ET, we use monthly estimates from the Global Land Evaporation Amsterdam Model (GLEAM) (Martens et al., 2017). We select GLEAM ET based on the findings of Koppa and Gebremichael (2017) in which GLEAM, AVHRR and MODIS ET datasets were ranked using a framework based on the Budyko hypothesis, a semi-empirical model that describes long-term water and energy balance of catchments (Budyko, 1974). The spatial resolution of the GLEAM dataset is 0.25° × 0.25°. We use monthly soil moisture estimates from ESA-CCI (Dorigo et al., 2017) for calibrating the hydrologic model with SM. The ESA-CCI dataset is selected based on the availability of data for the study period. The spatial resolution of the ESA-CCI dataset is 0.25° × 0.25°. We note that the ESA-CCI soil moisture measurements correspond to the top 5 cm of the soil layer, but the top soil layer of the hydrologic model (detailed below) is 10 cm. Although the unit of measurement is m^3/m^3, this difference in soil layer depths may lead to systematic bias and needs to be corrected. We adjust the values

Table 1
Noah-MP model physics options.

Model physics	Selected physics option
Vegetation model	Use table Leaf Area Index (4)
Canopy stomatal resistance	Ball-Berry (1) (Ball et al., 1987)
Soil moisture factor for stomatal resistance	Original Noah (1) (Chen et al., 1997)
Runoff and groundwater	TOPMODEL with groundwater (1) (Niu et al., 2007)
Surface layer drag coefficient	Original Noah (2) (Chen et al., 1997)
Frozen soil permeability	Linear effects, more permeable (1) (Niu and Yang, 2006)
Radiation transfer	Modified two-stream (1) (Yang and Friedl, 2003)
Snow surface albedo	CLASS (2) (Verseghy et al., 1991)
Rainfall and snowfall Partitioning	Jordan Scheme (1) (Jordan, 1991)
Lower boundary of soil temperature	Original Noah (2) (Chen et al., 1997)
Snow and soil temperature time scheme	Semi-implicit (1)
Super-cooled liquid water	No iteration (1) (Niu and Yang, 2006)

[a]The number in the brackets represents the internal Noah-MP model code for the selected physics option.

Table 2
Details of Noah-MP parameters used for calibration.

Parameter	Total parameters	Units	Minimum	Maximum
REFDK	1	m/s	1.4e-06	6.5e−06
REFKDT	1	No units	1.0	5.0
BB1 - BB12	12	No units	0.5	12.0
MAXSMC1 - MAXSMC12	12	No units	0.1	0.7
SATDK1 - SATDK12	12	m/s	2.0e−06	7.0e−02

[a]Soil texture classes for BB, MAXSMC and SATDK (from 1 to 12): sand, loamy sand, sandy loam, silt loam, silt, loam, sandy clay loam, silt clay loam, clay loam, sandy clay, silty clay and clay.

of simulated soil moisture to match the statistics of the observed dataset based on López López et al. (2017) as

$$SM'_{sim} = \frac{\sigma_{SM_{obs}}}{\sigma_{SM_{sim}}} * \left(SM_{sim} - \overline{SM}_{sim} + \overline{SM}_{obs} \right) \tag{5}$$

where SM'_{sim} is the scaled simulated soil moisture, $\sigma_{SM_{obs}}$ and $\sigma_{SM_{sim}}$ are the standard deviations of the observed and simulated soil moisture, SM_{sim} is the simulated soil moisture to be scaled, and $\overline{SM}_{obs}$ and $\overline{SM}_{sim}$ are the means of the observed and simulated soil moisture.

3.3. Setup of the hydrologic model

To replicate studies that use spatially distributed models, we choose the Noah-MP (Multi-Parameterization) Land Surface Model (LSM) (Niu et al., 2011), driven through NASA's Land Information System (LIS) (Kumar et al., 2006). The Noah-MP model builds on the original Noah LSM by incorporating a dynamic groundwater model, improved representation of vegetation canopy and snow pack. Cai et al. (2014) provide a detailed description and a comprehensive evaluation of the model over the Mississippi river basin. All the static input datasets required for running the Noah-MP model are sourced from NASA's LIS data portal (https://portal.nccs.nasa.gov/lisdata). The important static input datasets are the land cover map, sourced from USGS; the soil texture map from STATSGO, sourced from USDA; and the elevation map from GTOPO30, sourced from USGS. Albedo, greenness fraction and temperature are sourced from NCEP reanalysis. The meteorological forcings required by the Noah-MP model include precipitation, air temperature, surface pressure, specific humidity, wind speed, and radiation. All meteorological forcings are derived from Global Data Assimilation System (GDAS) from the Environmental Modeling Center (EMC) of the National Center for Environment Protection (NCEP) (Derber et al., 1991). The spatial resolution of the dataset is 0.47° × 0.47°. The meteorological inputs are interpolated onto the model grid using bilinear interpolation. To minimize the adverse effects of mismatch in the spatial resolution of the model and the reference datasets (Samaniego et al., 2010, 2017), the Noah-MP model is set-up for the Mississippi river basin at a spatial resolution of 0.25° × 0.25° (similar to the resolution of GLEAM ET and ESA-CCI SM). The Noah-MP model is spun-up for a period of 68 years by looping through the year 2003 until the groundwater and soil moisture storage reach equilibrium. The model time step is three hours. The number of soil layers in the model is four with thicknesses 10 cm, 30 cm, 60 cm and 100 cm. Specific Noah-MP model physics options selected for different processes are detailed in Table 1.

The Noah-MP model contains 71 standard parameters (present in user-defined Tables) and 139 hard-coded parameters (present in the model code). The Noah-MP model output has been found to be sensitive to about two-thirds of the 71 standard parameters (Cuntz et al., 2016). As the study is a calibration experiment involving multiple calibration cases, we keep the parameter dimension of the calibration problem manageable by selecting five of the most sensitive parameters from the Cuntz et al. (2016) study. The selected parameters are two surface runoff-related parameters (REFDK and REFKDT), the exponent in the Brooks-Corey equation (BB), soil porosity (MAXSMC), and hydraulic conductivity at saturation (SATDK). Of the five parameters, BB, MAXSMC, and SATDK are related to soil texture. Based on Cuntz et al. (2016), ET is sensitive to BB and MAXSMC parameters. SF (surface and subsurface runoff combined) is sensitive to all the five parameters with higher sensitivity to REFDK and REFKDT parameters. According to Cai et al. (2014), SM exhibits higher sensitivity to MAXSMC, SATDK. As there are twelve soil texture classes, the total number of parameters selected for calibration in the Noah-MP hydrologic model is 38 (Table 2 presents a detailed breakdown of the parameters with maximum and minimum values used for calibration). We select the minimum and maximum values of the parameters from literature (MAXSMC and SATDK values from Cai et al. (2014), BB and REFDK values from Cosby et al. (1984), and REFKDT values from Mendoza et al. (2015)). We subjectively adjust the minimum and maximum values to improve the rate of convergence of the calibration algorithms.

3.4. Setup of DREAM and AMALGAM algorithms

DREAM is a multi-chain Markov chain Monte Carlo (MCMC) simulation algorithm that automatically tunes the scale and orientation of the proposal distribution en route to the target distribution. It is designed for increasing the sampling efficiency of complex, high-dimensional parameter spaces, while maintaining detailed balance and ergodicity

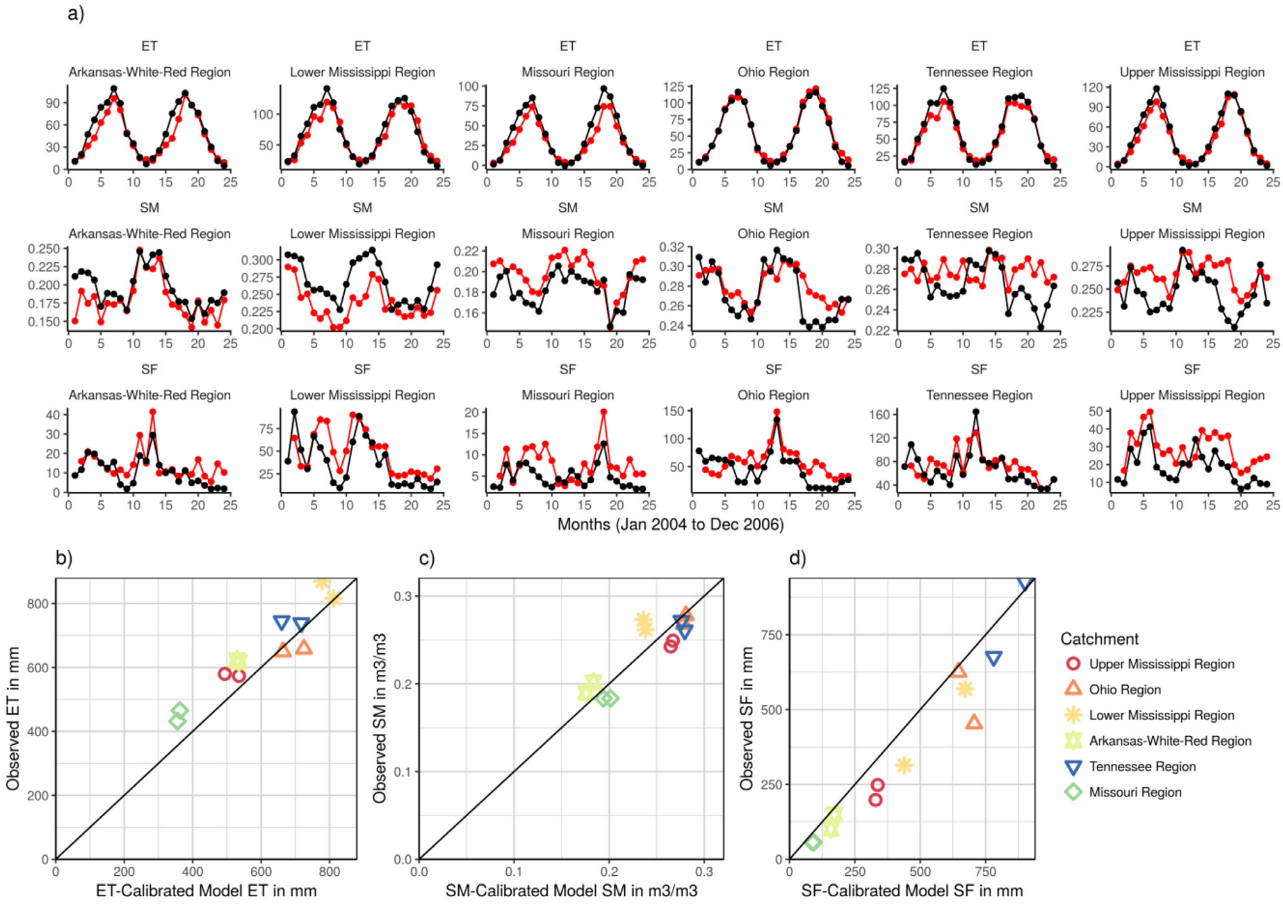

Fig. 4. Time series plots of (a) ET-calibrated model ET (red) and GLEAM ET (black) in mm/month (Top panel), (b) SM-calibrated model SM (red) and ESA-CCI SM (black) in m^3/m^3 (Middle panel) and (c) SF-calibrated model SF (red) and HUC-2 runoff (black) in mm/month (Bottom panel) for the six HUC-2 sub-catchments of the Mississippi basin. The first 12 months correspond to the calibration period of 2004 and the next 24 months correspond to validation year 2005; and (b) Scatter plot of annual modeled ET, SM, and SF and reference values for the calibration and validation years over the six HUC-2 basins. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

(Vrugt, 2016). In this study, we use the MT-DREAM (ZS) version of DREAM, which utilizes multi-try sampling (MT), snooker updating and sampling from an archive of past states to improve the rate of convergence and make use of parallel computing resources. Specific configuration options and parameters of the MT-DREAM (ZS) algorithm used in this study are detailed in Table 4. We select the Laplacian likelihood based on the findings of Schoups and Vrugt (2010); residual errors in rainfall-runoff models of humid basins, like the Mississippi basin (In Fig. 4d, most basins are within aridity index of 1.0), are better represented by a Laplacian distribution than a Gaussian distribution. The likelihood function is used to summarize the distance between the model simulations and the corresponding reference estimates. For ET and SM variables, error residuals determined at all 0.25°×0.25° grid cells and time steps (monthly) across the entire Mississippi river basin (all six HUC-2 basins together) are used to determine the likelihood function. Similarly, SF error residuals are determined using simulated and observed runoff at all six HUC-2 basins and all months of the calibration period. On a workstation with 16 processors, MT-DREAM (ZS) required approximately 16 days (14,500 iterations) to converge to a solution for each of the model responses (ET, SM, and SF), with each iteration of the Noah-MP model taking around 20 minutes to complete.

The multi-objective calibration algorithm, AMALGAM, combines the strengths of multiple evolutionary algorithms to improve the speed and efficiency of finding the Pareto optimal solutions for multi-objective optimization problems (Vrugt and Robinson, 2007). In the current implementation, four search algorithms are run simultaneously in AMALGAM: differential evolution (Storn and Price, 1997), particle swarm optimization (Kennedy and Eberhart, 2001), adaptive Metropolis (Haario et al., 2001), and NSGA-II (Deb et al., 2002). In AMALGAM, offspring creation is adaptive; the best performing algorithms in the present generation are weighted more in the creation of offspring for the next generation. Specific configuration options and parameter values of the AMALGAM algorithm used in this study are detailed in Table 3. As stated in the methodology section, the objective function is the root mean square error metric. For ET and SM, RMSE is estimated by calculating the error between the modeled and the observed quantities at each 0.25°×0.25° grid cell inside the entire Mississippi river basin and at each time step (each month of the year 2004). For SF, error residuals calculated for all six HUC-2 basins and twelve months are used for estimating RMSE. In other words, the calibration of the Noah-MP model is carried out for the entire Mississippi river basin using a single objective function (Laplacian likelihood in the case of DREAM and RMSE in the case of AMALGAM) and not for the individual HUC-2 basins. On a 16-processor workstation, each of the three multi-objective calibration scenarios (ET and SM, ET and SF, SM and SF) required around 10 days (9000 iterations) to arrive at the final Pareto front.

4. Results and discussion

First, we setup a synthetic numerical experiment to test whether the univariate calibration algorithm (DREAM) can uncover the true parameter values with one year of calibration data for ET, SM and SF

Table 3
Comparison of the true and calibrated parameters in the synthetic univariate calibration experiments.

Parameters	True	ET-calibration	SM-calibration	SF-calibration
REFDK	4.0e-06	4.17e−06 (1.48e−06)	4.45e−06 (1.52e−06)	4.12e−06 (1.19e−06)
REFKDT	3.0	2.88 (1.10)	3.38 (1.09)	3.12 (1.12)
BB	6	6.31 (3.34)	6.28 (3.17)	6.28 (3.28)
MAXSMC	0.4	0.42 (0.17)	0.41 (0.17)	0.41 (0.17)
SATDK	0.03	0.033 (0.021)	0.033 (0.027)	0.036 (0.025)

[a]Note: The values outside and inside the parenthesis represent the mean and standard deviation of the posterior distribution of calibrated parameters. 300 sample points of each parameter is used to calculate the mean and standard deviation values.

Table 4
MT-DREAM (ZS) and AMALGAM configuration.

DREAM otion	Specified option
Number of generations	600
Number of Markov chains	3
Number of forward model parameters	38
Number of crossover values	3
Number of Multi-tries	4
Number of chain pairs proposal	1
Likelihood function	Laplacian likelihood
AMALGAM Option	
Population size	150
Number of generations	60
Number of objective functions	2
Sampling strategy	Latin Hypercube

[a]Note: All other MT-DREAM (ZS) and AMALGAM parameters are set to default values.

variables. Next, we present the results of the Noah-MP model validation for the three model responses under consideration: evapotranspiration (ET), soil moisture (SM), and streamflow (SF). To help understand the impact of calibrating the Noah-MP model with a single water balance component (ET, SM, and SF) on other model responses and to define the limits of acceptability, we present the results of the posterior distributions of model errors (RMSE) from DREAM. We then test the central hypothesis of this study (Eq. 3) by combining the limits of acceptability and the multivariate calibration solutions from AMALGAM. Next, we address the second research question by analyzing the trade-offs among accurate simulations of ET, SM, and SF as represented by the Pareto fronts. Finally, we show how the developed multivariate calibration framework can help diagnose deficiencies in model parameterization.

4.1. Synthetic experiment

A synthetic numerical experiment is setup in the study domain (Fig. 3) to test whether the DREAM Bayesian calibration algorithm can successfully uncover the true parameter values that were used to generate the observations. The duration of the experiment is one year. First, we use the Noah-MP model to generate synthetic observations of ET, SM, and SF using a predetermined set of parameter values. For this, we select the same set of five parameters that are used for the real case study (Table 2). We use the mean of the parameter ranges as the ‘true’ value for each individual parameter (Table 3). For example, the true parameter value of REFDK is chosen as 4.0e-06. The Noah-MP model is run using the true parameter values and the meteorological forcing for the year 2004 to generate synthetic observations of ET, SM, and SF. These synthetic observations of ET, SM, and SF are then used to calibrate the Noah-MP hydrologic model without knowing the true parameter values. The configuration and computational budget of the DREAM algorithm is consistent with that of the real case study (Table 4). We compare the resulting parameter distributions from the three calibration cases (ET, SM, and SF) to test whether the calibrated values converge to the true parameter values. Table 3 presents the results of the synthetic numerical experiment. From Table 3, it is evident that the calibration of the Noah-MP model using the synthetic observations converges to the true values for all three cases (ET, SM, and SF). For the streamflow related parameters REFDK and REFKDT, calibration with synthetic ET and SF observations produced the best convergence in terms of mean and standard deviation of the posterior distribution. Of the other parameters, the highest variance is seen in the parameter BB. Although the mean of the posterior distributions are close to the true parameter value, the standard deviation is high. All the calibration cases (ET, SM, and SF) converge to the true parameter values for MAXSMC and SATDK values. In the subsequent sections, we present the results of the application of the calibration framework for the Mississippi river basin.

4.2. Validation of the Noah-MP hydrologic model

To ascertain whether the Noah-MP model and the parameters considered for calibration can simulate ET, SM, and SF accurately, we evaluate the model for the year 2005. For this, we select a parameter set from the DREAM solutions that results in the lowest RMSE value for each of the model responses (ET, SM, and SF), thus having three parameter sets, each representing a benchmark for ET, SM and SF variables. We present a time series comparison of observed and simulated monthly ET, SM, and SF for the six HUC-2 sub catchments (Fig. 4a).

When the Noah-MP model is calibrated with GLEAM ET (top panel), it is evident that the model performs very well in simulating the observed ET values and seasonality. One exception is the underestimation of ET in the summer months for all regions except the Ohio region. The results are consistent across the six HUC-2 hydrologic regions in the calibration (first 12 data points) and validation time periods (remaining 12 data points). The close match between modeled and observed ET is reflected in the scatter plots of the annual totals of ET as well (Fig. 4b).

Relatively higher variance is observed in the SM simulations when the Noah-MP model is calibrated with ESA-CCI soil moisture (middle panel). The simulated soil moisture for all the regions is quantitatively consistent with the observed SM values for the top soil layer. However, we see some discrepancy in the seasonality of soil moisture between the simulated and observed soil moisture; it is especially pronounced in the last six months of 2005 in the Tennessee region (middle panel, fifth column) and the first few months of 2004 in the Arkansas-White-Region (middle panel, first column). In the Missouri region (middle panel, third column), the model simulates the observed seasonality but is unable to capture the peaks in the observed SM perfectly. In the Upper Mississippi region (middle panel, sixth column), the model has difficulty in simulating the timing of the troughs (May and June of 2004 and 2005) seen in the observed SM. Similar to the ET results, the annual average SM consistently matches the observed values for all six HUC-2 basins (Fig. 4b).

The results for streamflow simulated by the SF-calibrated Noah-MP model present a more consistent picture (bottom panel). They show that the model generally performs well for all six HUC-2 catchments. However, similar to soil moisture results, there are some inconsistencies in simulating the seasonality of the last six months of 2005, especially in the Arkansas-White-Red (bottom panel, first column), the

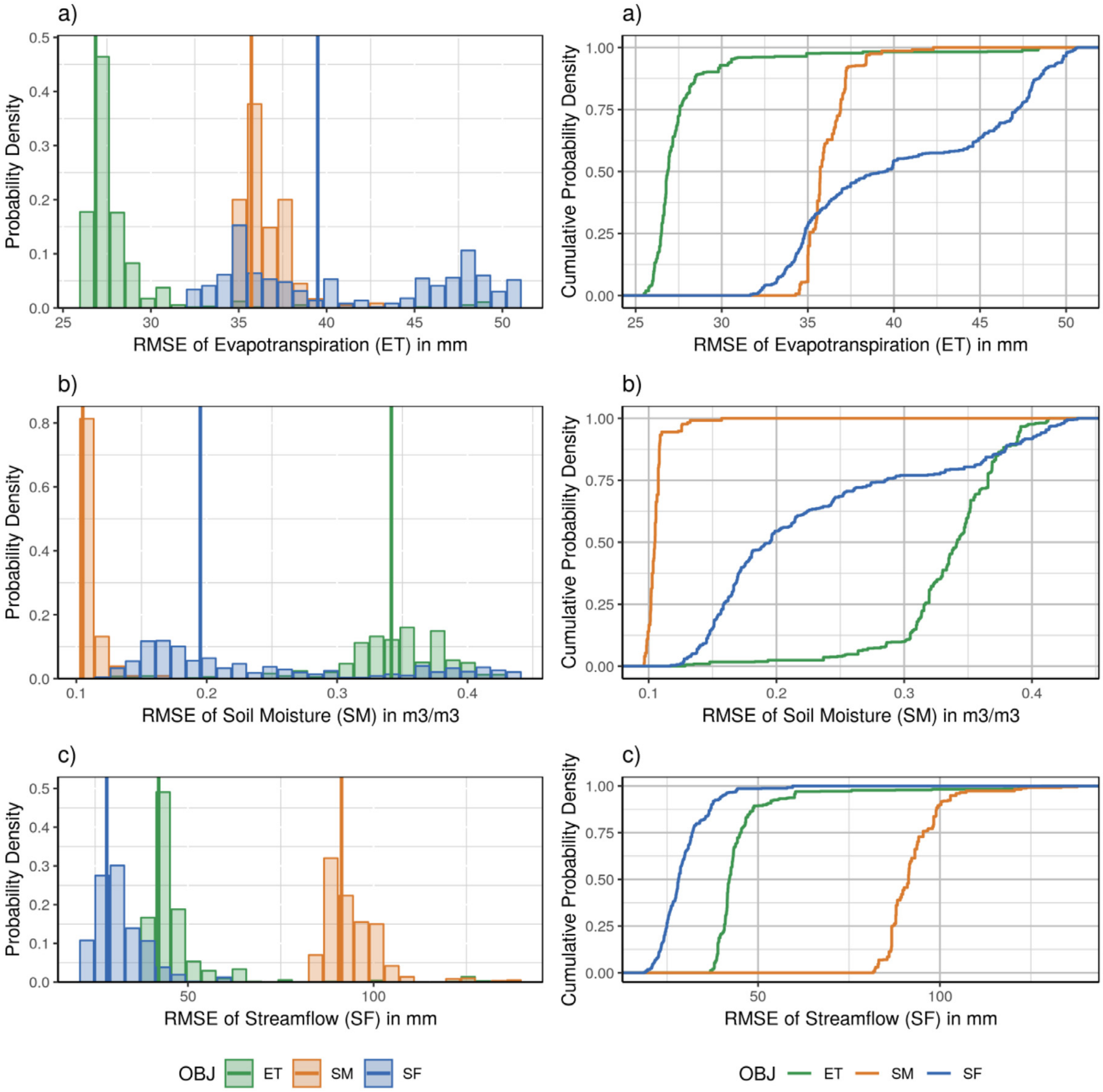

Fig. 5. A comparison of the posterior probability density functions (PDF) and empirical cumulative distribution functions (ECDF) of root mean square errors of (a) evapotranspiration (top panel), (b) soil moisture (middle panel) and (c) streamflow (bottom panel) when the model is calibrated using DREAM with ET (green), SM (orange) and SF (blue). Vertical lines in the PDFs represent 50% quantiles of RMSE. For comparison, the mean annual water balance closure error from the reference datasets (P - Q - ET) is around 108 mm (9 mm/month). Note: (1) The RMSE values of ET and SM are determined from error residuals calculated at all grid cells of the model domain. The RMSE values of SF are determined using error residuals from all the six HUC-2 regions. (2) 300 sample points of model error (RMSE) are used to construct each PDF and ECDF. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

Ohio (bottom panel, fourth column) and the Upper Mississippi (bottom panel, sixth column) regions. The model also performs well in the Lower Mississippi and Missouri regions, but the peaks are higher in 2004 (June and July) compared with the observed streamflow time series. At annual timescales, the SF-calibrated Noah-MP model overestimates SF for the year 2005 in the Ohio basin (June to October 2005). We see that the errors in evapotranspiration and streamflow are comparable to the results of the Ma et al. (2017) study. For the six HUC-2 subbasins considered in this study, Ma et al. (2017) report a RMSE of about 10 mm/month, whereas the RMSE in this study is about 18 mm/month. The average RMSE of ET, calculated over the entire US by *Ma et al.* (2017), is about 10 mm/month, which matches the RMSE of the time series of ET presented in Fig. 4 (about 10 mm/month). We note that the Ma et al. (2017) study does not calibrate the Noah-MP model. In addition, the Ma et al. (2017) uses FLUXNET ET data as reference observations for calculating the RMSE.

4.3. Posterior distributions of model errors

We present the posterior distribution of model response errors (RMSE) using probability density and empirical cumulative distribution functions (PDF and ECDF) (Fig. 5). The PDFs and CDFs of three model responses (ET (top panel), SM (middle panel) and SF (bottom panel)) are constructed for three calibration objectives: ET (green), SM (orange) and SF (blue). First, we analyze the impact of calibrating the Noah-MP hydrologic model with only streamflow (blue) on ET (top panel) and SM (middle panel). Our results confirm the findings of previous studies: calibrating a hydrologic model with only SF adversely affects the accuracy of other model responses (Rakovec et al., 2016a,b). While the 50% quantile of the ET error increases by about 15 mm/month, the SM error increases by around 0.1 m^3/m^3. The same conclusion can be extended for univariate calibration with other variables (ET and SM). For example, when the Noah-MP model is calibrated with SM (orange), the

Table 5
Limits of acceptability for ET, SM, and SF derived from posterior distributions of RMSE.

Quantile	ET (in mm)	SM (in m^3/m^3)	SF (in mm)
10%	26.1	0.098	22.7
25%	26.5	0.101	24.9
50%	26.9	0.104	28.1
75%	27.6	0.107	32.0
90%	29.0	0.108	37.30
95%	30.6	0.126	40.42
99%	47.9	0.132	58.00

Table 6
Breakdown of Pareto optimal solutions into behavioral solutions based on different limits of acceptability for ET-SM, ET-SF and SM-SF combinations.

Quantile	ET and SM ($N = 123$)	ET and SF ($N = 29$)	SM and SF ($N = 74$)
10%	0	0	0
25%	0	0	0
50%	0	12	0
75%	0	25	0
90%	0	29	12
95%	12	29	27
99%	86	29	58

[a] N represents the total number of Pareto optimal solutions derived from AMALGAM using an initial population of 150.

errors in ET (top panel) and SF (bottom panel) are very high. Some results stand-out from Fig. 5: (1) The error distribution of SF produced with the ET-calibrated model is close to the error distribution of SF from the SF-calibrated model (green and blue plots in bottom panel). This finding reinforces the results of previous studies that use ET for calibration (Immerzeel and Droogers, 2008; López López et al., 2017; Zink et al., 2018). However, the absolute difference between the 50% quantiles is still around 20 mm/month, indicating a significant impact on SF accuracy due to univariate calibration of the model with ET; 2) Similarly, SF-calibration seems to produce lower errors for SM compared to ET (middle panel) but the absolute errors are still very high; 3) Neither SM nor SF calibration can simulate ET with reasonable accuracy, as seen by the disparate error distributions (top panel). In fact, both SM and SF produce similar error distributions for ET, evident by the ECDFs (blue and orange). For comparison, the RMSE of ET and SF are higher than the mean monthly water balance closure error seen in the reference dataset (about 9 mm/month). Therefore, the RMSE values of ET and SF are not significantly biased by the errors in the reference datasets, as represented by the water balance closure error. We derive the limits of acceptability for ET, SM, and SF from the posterior distributions of RMSE. The error thresholds are defined at 10%, 25%, 50%, 75%, 90%, 95% and 99% quantiles (Table 5). The table shows that the range of the quantiles is quite low for ET (about 4 mm/month between the 10% and 95% quantiles) and SM (about 0.02 m^3/m^3 between the 10% and 95% quantiles). This is reflected in the well-defined posterior distributions of ET and SM errors when the model is calibrated with ET (Fig. 5, green, top panel) and SM (Fig. 5, orange, middle panel). In the case of SF, the difference between the 10% and 95% quantiles is about 17 mm/month. This discrepancy maybe due to the higher number of data points available for calibrating ET and SM (spatially distributed reference datasets) compared to SF (point data).

4.4. Hypothesis testing using DREAM and AMALGAM solutions

To address the first research question, we test the hypothesis that the intersection of behavioral (from DREAM) and non-dominated solutions (from AMALGAM) is a non-empty set (Eq. 3). We present the Pareto fronts of the RMSE of different combinations of model responses (ET-SM, ET-SF, SM-SF) along with different limits of acceptability (Table 6) in Fig. 6. As stated in the methodology section, multivariate calibration with evolutionary algorithms such as AMALGAM may result in Pareto fronts with lesser numbers of non-dominated solutions compared with the initial population size. In this study, we focus only on the Pareto optimal solutions (red stars in Fig. 6). The non-dominated or Pareto optimal solutions are also used to analyze the Pareto fronts and quantify the trade-offs (next section). We present the number of Pareto optimal solutions of different combinations of model responses that lie within specific RMSE thresholds in Table 6.

For the combination of ET and SM model responses (Fig. 6a), we see that the hypothesis (Eq. 3) fails for all defined error thresholds except for the 95% and 99% quantiles (Table 6 for specific values). Even at the 95% quantile, where the ET and SM error thresholds are quite high (30.6 mm/month for ET and 0.126 m^3/m^3 for SM), only 12 points in the multivariate space can be classified as behavioral. We note that the points near the tails of the Pareto front are well within individual error thresholds of ET and SM. Therefore, we conclude that the Noah-MP model is unable to simultaneously simulate both ET and SM with reasonable accuracy. Even though the incorporation of ET and SM can improve the errors in both the variables, it does not result in improving the realism of the model itself, lending credence to assertion that the relationship between ET and SM may not be complementary. The results are consistent with those of univariate calibration using DREAM, wherein the distributions of the ET- and SM- calibrated model responses have considerable discrepancies between them (compare ET-calibrated ET and SM-calibrated ET (top panel of Fig. 5) and SM-calibrated SM and ET-calibrated SM (middle panel of Fig. 5)).

In contrast to ET and SM, the model performs very well in simulating both ET and SF accurately (Fig. 6b). Even at a stricter error threshold of 50% quantile of individual ET and SF errors (Table 5), around 12 points out of the 29 Pareto optimal points can be classified as behavioral (Table 6). At the error threshold of the 95% quantile, all 29 Pareto optimal solutions are behavioral. This, along with the fact that the range of errors in the Pareto front is quite small, shows that the relationship between ET and SF can be considered complementary. In other words, the incorporation of ET into SF calibration can improve ET simulation while maintaining the accuracy of SF within reasonable error thresholds. As seen in Fig. 5, the presence of a complementary relationship between ET and SF is hinted in the posterior distributions (PDF and ECDF) of ET and SF errors (SF-calibrated SF (blue) and ET-calibrated SF (green) in the bottom panel). This result also provides support for studies that incorporate ET and SF into calibration, such as Zink et al. (2018), where incorporation of ET improved ET error by 8% while the NSE of SF reduced by 6% (which could be within the limits of acceptability).

For the combination of SM and SF flux, the hypothesis fails for the lower error thresholds (10%–50% quantiles). At higher quantiles (75% and greater), where the absolute value of the SF error threshold is greater than 32 mm/month (Table 5), relatively more Pareto optimal solutions are classified as behavioral compared to the ET-SM combination. This relative improvement in performance also is reflected in the posterior distribution of individual model response errors (middle panel in Fig. 5). We see that the error distribution of SF-calibrated SM (blue) is closer to the SM-calibrated SM (orange), compared with the ET-calibrated SM (green). This, combined with the results of the ET-SM Pareto front, shows that incorporation of a storage component such as SM may not lead to improved model performance for multiple water balance components compared with the incorporation of ET. The performance of the SF-SM combination is in line with the findings of Wanders et al. (2014), where there are improvements in SF and SM performance when SM is incorporated into calibration. This also highlights the advantages of defining limits of acceptability and casting multivariate calibration as a trade-off problem. Similarly, the results show the drawbacks of assuming a complementary relationship between the

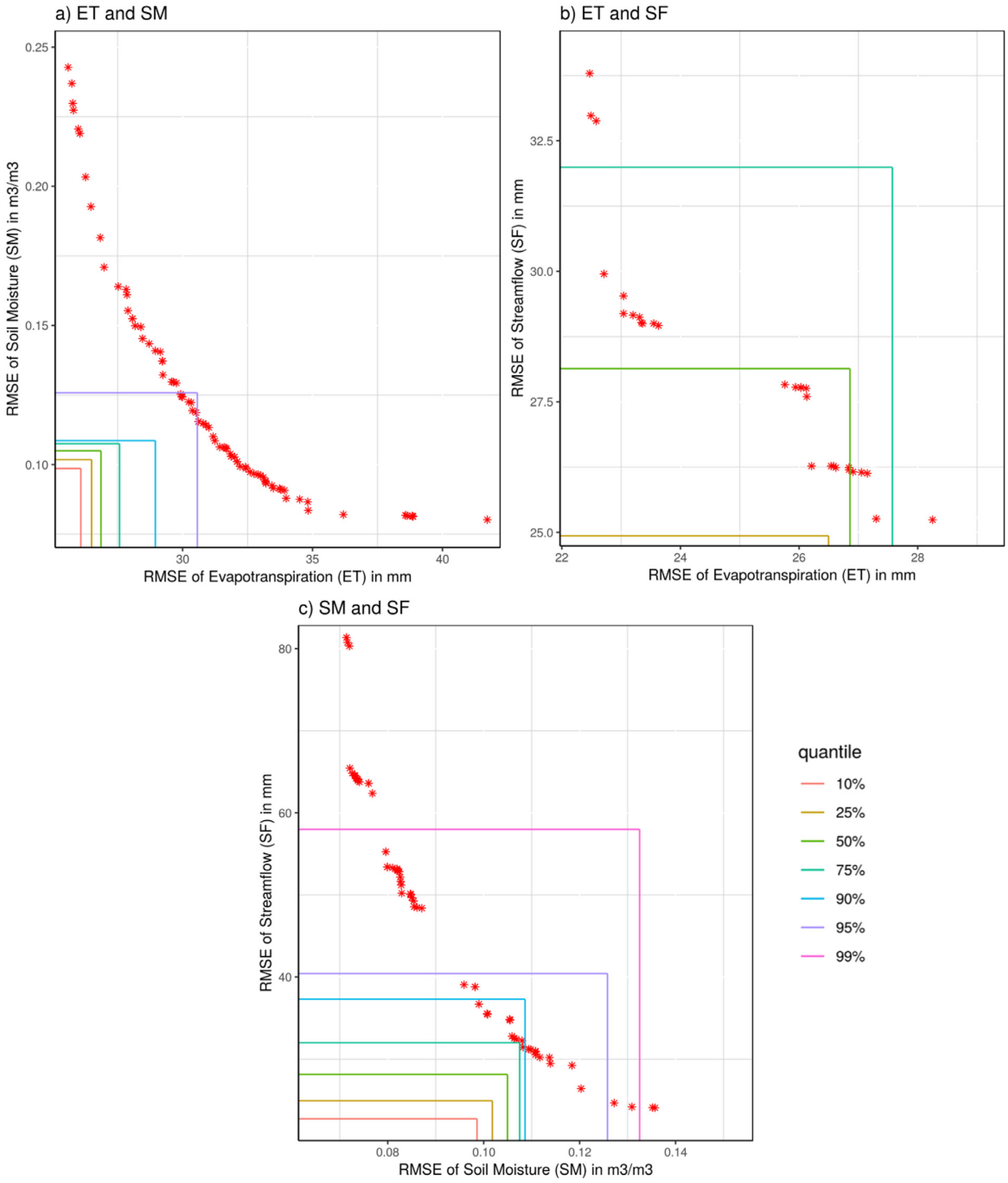

Fig. 6. Pareto fronts of root mean square errors of (a) ET and SM, (b) ET and SF and (c) SM and SF with limits of acceptability represented by 10%, 25%, 50%, 75%, 90%, 95% and 99% quantiles of the posterior distribution of the error (from DREAM). For comparison, the mean annual water balance closure error from the reference datasets (P - Q - ET) is around 108 mm (9 mm/month). The non-dominated or Pareto optimal solutions are represented by red stars. Note: The RMSE values of ET and SM are determined from error residuals calculated at all grid cells of the model domain. The RMSE values of SF are determined using error residuals from all the six HUC-2 regions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

model responses. For example, in Rakovec et al. (2016b), the incorporation of TWS improves the RMSE (normalized) of TWS from 0.9 to 0.8 with low impact to SF performance. However, without the specification of an error threshold, it cannot be determined whether the RMSE of 0.8 is behavioral. In other words, the solution may lie on the Pareto front but may still be outside the limits of acceptability. Therefore, the realism of the model may not have improved to a sufficient degree due to the incorporation of an additional water balance component.

4.5. Understanding the trade-offs using Pareto fronts

The results of the hypothesis tests could be a consequence of the nature of trade-offs between the objectives of multivariate calibration,

Table 7
Average, minimum and maximum trade-offs for combinations of ET, SM, and SF model responses (minimum and maximum values are within parentheses).

	ET	SM	SF
ET	–	2.0 (0.5, 7.5)	0.3 (0.08, 0.9)
SM	0.5 (0.1, 2.0)	–	1.0 (0.24, 2.9)
SF	2.0 (1.1, 12)	0.7 (0.35, 4.2)	–

as represented by the Pareto fronts (Fig. 6). In this section, we address research question 2 and diagnose the results of hypothesis testing. We analyze the characteristics of the trade-offs in accurately simulating the three combinations of model responses incorporated into calibration - 1) ET-SM, 2) ET-SF, and 3) SM-SF. A visual analysis of the Pareto fronts (Fig. 6) reveals that the ET-SM and SM-SF fronts are relatively well defined compared to the ET-SF Pareto front. The ill-defined ET-SF Pareto front could be an indication that the relationship between ET and SF is complementary, as opposed to a trade-off relationship. We compare the number of non-dominated solutions in the multivariate space to test whether these numbers reflect the conclusions of the visual analysis. We see that the ET-SM front has the highest number of non-dominated solutions (123 solutions), followed by SM-SF (74 solutions), and then ET-SF (29 solutions). The higher number of non-dominated or Pareto optimal solutions in the ET-SM front can also indicate a strong trade-off relationship between the ET and SM model responses compared with other combinations. This qualitatively seems to confirm the conclusions drawn from testing the central hypothesis of this study. First, it is consistently more difficult to accurately simulate ET and SM together compared with other combinations. On the other end of the spectrum, the ET and SF fluxes are more complementary to each other, considering 1) the ill-defined shape of the Pareto front and 2) the lesser number (29) of Pareto optimal solutions. In the case of SM and SF model responses (Fig. 6c), the higher accuracy of the SF-calibrated hydrologic model in simulating SM compared with an ET-calibrated model (middle panel in Fig. 5) translates to a lesser number of non-dominated solutions compared with the ET and SM combination.

However, the number of Pareto optimal solutions is not a quantitative measure of the trade-off among the water balance components, as they depend on the optimization algorithm used. Next, we analyze the accuracy trade-offs in simulating the combination of model responses quantitatively. First, we compare the range of the objective functions in each of the three multivariate calibration cases. In the case of ET and SF, the range of the two objectives is very small compared with other combinations. For example, the RMSE of ET flux ranges from 25.5 mm/month to 41.7 mm/month in the ET-SM combination, whereas in the ET-SF combination the range is between 22.5 mm/month and 28.9 mm/month. Similarly, the range of SF in the ET-SF combination (25.2–34.1 mm/month) is lower than in the SM-SF combination (24.1–81.3 mm/month). Also, the SM range is lower in the SM-SF combination (0.07 to 0.14 m^3/m^3) compared to the ET-SM combination (0.08–0.24 m^3/m^3). Lower ranges of the objectives in the ET-SF combination imply that both ET and SF can be simulated together to a reasonable degree of accuracy. Lower ranges also may point toward a lower trade-off between two objectives. However, due to differences in the units of the objective functions and different ranges across the three Pareto fronts, more analysis is required to draw conclusions.

Finally, we compare the magnitude of trade-offs (defined in the methodology section) across the Pareto fronts of the three calibration cases (ET-SM, ET-SF, and SM-SF). To enable comparison of the trade-offs across the three Pareto-fronts, we normalize the errors in each of the model responses (ET, SM, and SF) by the maximum error, determined over all three combinations. To derive the trade-off matrix (Table 7), we fit a second-degree polynomial to the non-dominated points (red stars in Fig. 6) of the three combinations of model responses - ET-SM ($R^2 = 0.92$), ET-SF ($R^2 = 0.82$), and SM-SF ($R^2 = 0.96$). Next, we divide the Pareto fronts into 15 equally spaced segments. We then calculate the average, maximum and minimum trade-offs from either the slopes or the inverse of the slopes of the 15 segments for each combination of the three model responses. For the combination of ET and SM (first row, second column in Table 7), the average increase in the error of ET required to affect a unit decrease in the error of SM is 2.0 units. For the same improvement, the SF error trade-off is only 0.7 units, implying a larger trade-off in the ET-SM combination compared with the SM-SF combination for soil moisture. Confirming this conclusion, we see that the average trade-off in ET accuracy required to improve SF is much lower than the trade-off required to improve SM. It is interesting to note that there is a higher trade-off in SF accuracy to achieve a unit improvement in ET compared to the trade-off in SM accuracy to achieve the same improvement in ET (first column in Table 7). However, as Fig. 6 shows, more ET-SF multivariate calibration solutions are behavioral compared to ET-SM solutions, and the range of errors is much lower in the ET-SF combination.

4.6. Model diagnosis

Unlike deterministic calibration where a single optimal parameter set is derived, the multivariate calibration framework developed in this study enables the examination of parameter behavior among multiple calibration objectives. Hogue et al. (2006) demonstrated the advantages of Pareto calibration for studying model performance and parameter behavior. In this study, we explore the advantages of combining behavioral solutions from Bayesian calibration and Pareto optimal solutions from multivariate calibration for model diagnosis. The objective is to identify parameters that can explain the significant trade-offs in accuracy among the multiple fluxes and storage variable incorporated into calibration. Specifically, we try to investigate the reasons behind the higher magnitude of trade-offs in the ET-SM and ET-SF cases. We first compare the posterior probability distributions of calibrated parameters for the univariate calibration cases (ET, SM, and SF) to identify the parameters that govern the performance of the Noah-MP model with respect to the ET, SM, and SF responses. Fig. 7 presents a comparison of the ECDFs of the five parameters considered for calibration in this study - REFDK, REFKDT, BB, MAXSMC and SATDK. We note that the ECDFs for parameters BB, MAXSMC and SATDK have been constructed by aggregating parameter solutions from all 12 soil classes (Table 2). To keep the analysis consistent with multivariate calibration results (presented later), we only consider behavioral solutions that are within a threshold of the 50% quantile. The ECDFs of the runoff parameters, REFDK and REFKDT, explain the performance of the Noah-MP model with respect to streamflow for different univariate calibration cases. We see that the ECDFs of both REFDK and REFKDT for the SM-calibrated model (orange) are quite different from the SF-calibrated model (blue), leading to relatively poor performance (Fig. 5). We see that the REFDK parameter distribution from the ET-calibrated model (green) is closer to the REFDK parameter distribution from the SF-calibrated model (blue) compared to the SM-calibrated model (orange). On the other hand, the REFKDT parameter distributions from the ET-calibrated and SM-calibrated models are considerably different from the SF-calibrated model. This shows that the relatively better performance of the ET-calibrated model for the SF flux compared to the SM-calibrated model is more influenced by the REFDK parameter.

We quantify these differences between the parameter PDFs using the Hellinger's distance for PDFs (Eq. 4), and the distance between their ECDFs using the Kolmogorov-Smirnov (KS) test statistic. The distance measures are presented in the form of a heat map in which each grid cell represents the statistical distance between the corresponding parameter distributions. For example, the first grid cell in the first panel of Fig. 8a represents the Hellinger's distance between the PDFs of the REFDK parameter generated from the ET-calibrated model (first column) and the SF-calibrated model (first row). The three rows in each

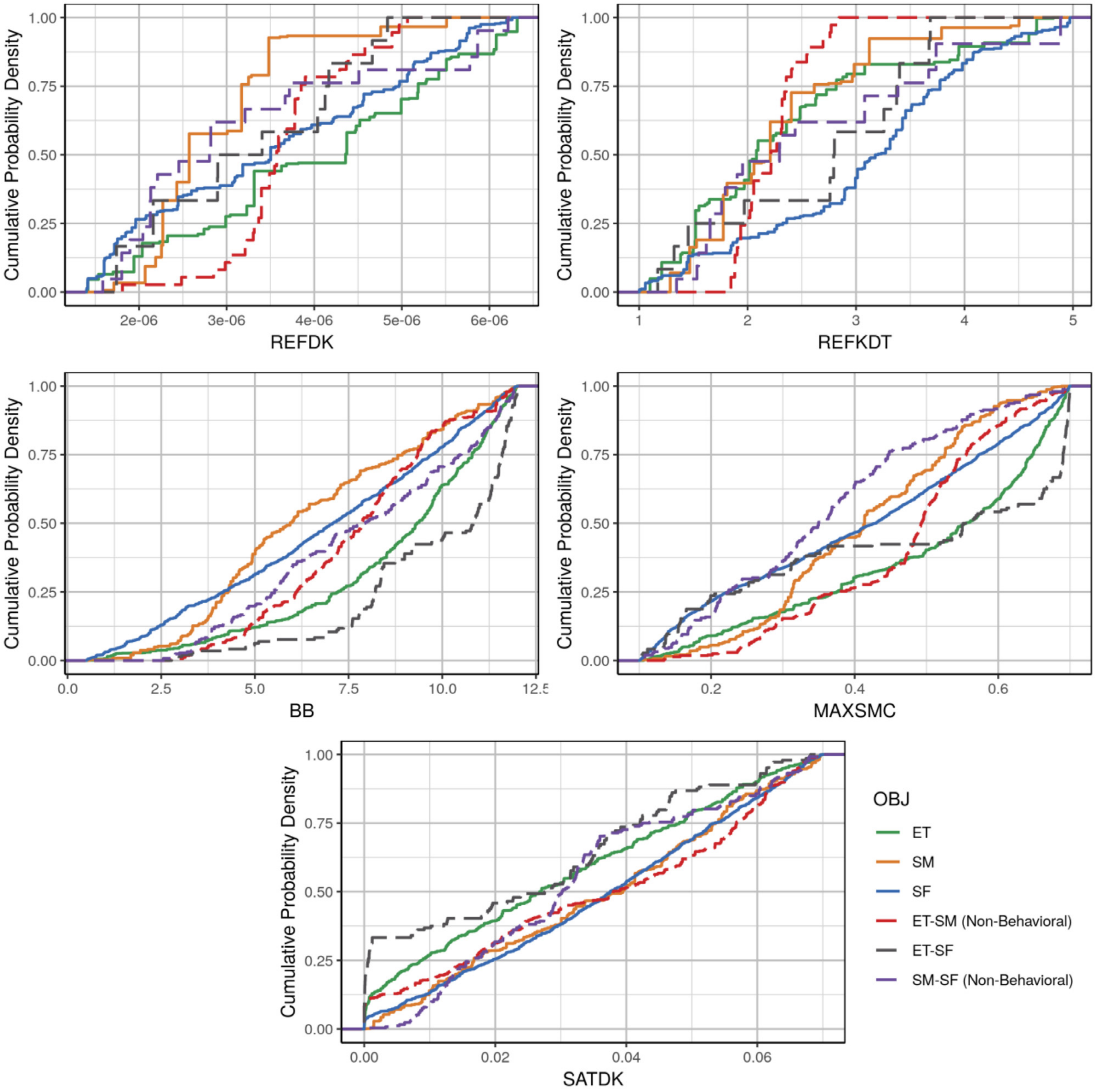

Fig. 7. Empirical cumulative density functions (ECDF) of calibrated Noah-MP parameters for univariate (ET, SM, and SF) and multivariate (ET-SM, ET-SF, and SM-SF) objectives. For the ET, SM, SF, and ET-SF objectives, the solutions within the error threshold of the 50% quantile are used to construct the PDFs and ECDFs of the parameters. For the ET-SM and SM-SF solutions that are not behavioral for both, model responses at the 50% quantile error threshold are used. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

panel correspond to the univariate calibration objectives (ET, SM, and SF), and the six columns correspond to both the univariate and multivariate calibration objectives (ET-SM, ET-SF, and SM-SF). We see that the Hellinger's distance between the REFDK distributions generated from the ET-calibrated model and the SF-calibrated model (third column and third row) is less than the distance between the distributions generated from the SM-calibrated model and the SF-calibrated model (third column and second row). This difference is more pronounced in the KS statistic heat map (Fig. 8b). As far as ET is concerned, the inability of both the SM- and SF-calibrated models to accurately simulate ET can be attributed to differences in the posterior distributions of the BB and MAXSMC parameters. As pointed out by Cuntz et al. (2016), the runoff parameters also can influence the ET (and SM) results, as they affect the model's water balance. This fact is evident in the case of SM performance, for which the parameter distributions from both the SM-calibrated and SF-calibrated models for BB, MAXSMC and SATDK are similar. However, there is a large difference in the distributions of the runoff parameters, REFDK and REFKDT (Fig. 8).

The analysis of posterior distributions of parameters from univariate calibration presented above agrees with the results of sensitivity analysis of the Noah-MP model parameters (Cuntz et al., 2016). As land surface models are highly complex in terms of parameterization, the same conclusions may not hold when multiple fluxes and state variables are incorporated into calibration. We calculate the correlation between the objective functions (RMSE) for the combinations of model responses (ET-SM, ET-SF, and SM-SF) and the corresponding parameters (REFDK, REKDT,

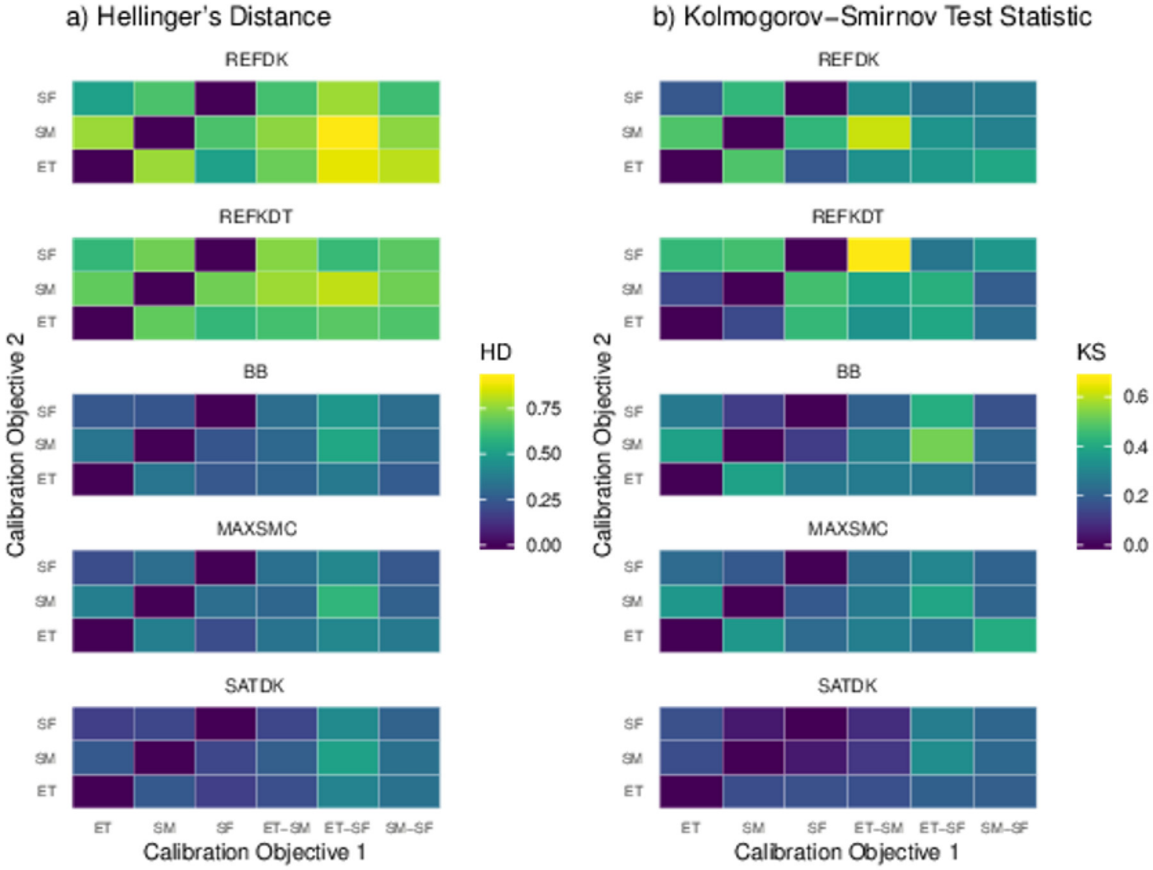

Fig. 8. (a) Hellinger's distance between PDFs of the calibration parameters for different calibration objectives and (b) the Kolmogorov–Smirnov test statistic between the ECDFs of calibration parameters for different calibration objectives. For the ET, SM, SF and ET-SF objectives, the solutions within the error threshold of the 50% quantile are used to construct the PDFs and ECDFs of the parameters. For ET-SM and SM-SF solutions that are not behavioral for both, the model responses at the 50% quantile error threshold are used.

Table 8
Correlation between objective functions (RMSE) and Noah-MP model parameters (REFDK, REFKDT, BB, MAXSMC and SATDK) from the Pareto optimal solutions for three multivariate calibration cases (ET-SM, ET-SF, and SM-SF).

Objectives	REFDK	REFKDT	BB	MAXSMC	SATDK
ET and SM	−0.57, 0.62)	(0.50, −0.16)	(−0.37, 0.27)	(−0.32, 0.17)	(0.12, −0.05)
ET and SF	(−0.27, 0.24)	(0.09, −0.02)	(−0.17, 0.17)	(0.04, −0.07)	(−0.17, 0.18)
SM and SF	(−0.15, 0.13)	(−0.02, 0.12)	(0.24, −0.28)	(−0.21, 0.20)	(0.13, −0.10)

[a] The numbers in the parenthesis represent correlation of the parameters with objective functions 1 (ET in first row) and 2 (SM in first row), respectively.
[b] For the BB, MAXSMC and SATDK parameters the median correlation from the 12 soil classes are presented.

BB, MAXSMC, and SATDK). In other words, we map the behavior of the parameters along the Pareto front. From Table 8, it is clear that most parameters show opposite correlation for the pair of objectives. For the combination of ET and SM objectives, parameters REFDK and BB show strong correlation with both ET (positive) and SM (negative) errors. This may indicate that these parameters are more responsible for the higher trade-off in accuracy (slope of the Pareto curve) between ET and SM seen in Fig. 6. Most parameters have almost equal magnitude (but different signs) of correlation with the objective functions, except for REFKDT. In the ET-SF combination, it is clear that the magnitude of correlation between the objective functions and the parameters are significantly lower than the correlations for the ET-SM combination. This clearly reflects the lesser trade-off and better simultaneous simulations of ET and SF compared with ET and SM. It is interesting to note the higher correlation for the REFDK parameter compared with REFKDT, considering that the difference between the ET-calibrated and SF-calibrated REFKDT parameters was higher than the REFDK parameter. This shows that the parameters that govern model performance for individual water balance component may not correspond to the parameters that affect the trade-off among the combination of water balance components. Also, it highlights the advantage of employing Pareto-based calibration. A similar conclusion can be drawn when SM and SF are combined; we see that the trade-offs are governed by BB and MAXSMC parameters and not the REFDK or REFKDT variables that govern the performance of SF.

Next, we compare the parameter distributions of behavioral multivariate calibration solutions of the ET-SF combination with parameter distributions of models calibrated with only ET and SF. Behavioral solutions with respect to both incorporated model responses are derived at the 50% quantile threshold (Table 5). As the number of ET-SF Pareto optimal solutions within the 50% behavioral threshold are few (12), we use both the Pareto optimal and dominated solutions (40) to derive the ECDFs of the five parameters. We note that the forty Pareto optimal and dominated solutions used for deriving the parameter distributions are behavioral (within 50% error quantile), and therefore can be considered as valid multivariate calibration solutions. In addition, we stress here that, unlike DREAM, the parameter distributions derived from multivariate calibration are not true posterior distributions. Therefore, we intend to use the analysis of parameter distributions from multivariate calibration only as a tool to investigate difference in parameter values between univariate and multivariate calibration solutions. First, we compare the distance between behavioral ET-SF and SF-calibrated solutions. We see that the parameter distribution of the REFKDT parameter in the ET-SF calibrated model is much closer to the REFKDT distribution in the SF-calibrated model (fifth column and first row in Fig. 8) compared with the REFKDT distribution in the ET-calibrated model (fifth column and third row in Fig. 8). However, the Hellinger's distance and KS statistic between the ET-calibrated and SF-calibrated models (first column and first row in Fig. 8) for the REFDK parameter is smaller compared to the distance between behavioral ET-SF solutions and the SF-calibrated model (fifth column and first row in Fig. 8). In fact, the incorporation of ET and SF into the calibration only improves the REFKDT parameter distribution, and all other parameters show greater distances compared with parameters derived from models calibrated with only ET or SF. This shows that the surface runoff parameter, REFKDT, is very dominant in governing the behavioral simulation of both the ET and SF fluxes. Finally, we try to determine the reasons for the poor combined simulation of the ET-SM and SM-SF variables. To do this we compare the multivariate calibration solutions from the ET-SM and SM-SF calibration cases that are not behavioral with respect to any of the individual fluxes or storage variable in models calibrated with only the individual fluxes or storage variable (ET, SM, and SF). Considering the case of the ET-SM combination, we see that the incorporation of both the ET and SM responses reduce the distances between the distributions of parameters that govern ET (fourth column and third row for BB, MAXSMC and SATDK in Fig. 8) compared with the SM-calibrated model (fourth column and second row for BB, MAXSMC and SATDK in Fig. 8). However, the distributions of the REFDK variable, which seem to influence the trade-off in accuracy between the ET and SM variable (Table 8), deviate significantly from the SF-calibrated model (compare Hellinger's distance between the ET-SM combination and SF, and the ET-SM combination and ET for the REFDK parameter in Fig. 8). This shows that a surface runoff parameter such as REFDK can influence the combined simulation of ET and SM. We can draw a similar conclusion in the case of SM-SF, where the incorporation of SM and SF improves the performance of the runoff parameters (REFDK and REFKDT) compared with the SM-calibrated model. However, the parameter that seems to influence the trade-off more, MAXSMC, shows degradation compared with the SM-calibrated model. It is interesting to note that the combination of SM and SF adversely affects a soil moisture parameter, even though SM has been incorporated.

5. Conclusions and future work

With the rise in availability of satellite-based measurements of hydrologic fluxes and states (Lettenmaier et al., 2015), multivariate calibration is widely seen as a promising solution for improving the performance and realism of large scale hydrologic models. However, most multivariate calibration studies do not formally define any acceptable error thresholds to help one conclude whether incorporation of additional water balance components into calibration improves either the performance or the realism of a hydrologic model. In addition, apriori assumptions such as complementary relationships between the different water balance components and deterministic calibration approaches hinder rigorous testing and diagnosis of hydrologic models, as called-for by Beven (2018). In this study, we develop a framework for multivariate calibration by combining Bayesian and Pareto optimality-based calibration methodologies. This framework can be used to test whether models simultaneously can simulate multiple water balance components accurately by accepting or rejecting parameter solutions based on a defined error threshold. Applying the framework to a large scale distributed hydrologic model (Noah-MP), we find that the model simulates different combinations of different fluxes and a storage variable (ET and SM, ET and SF, and SM and SF) with varying degrees of acceptability. While ET and SF can be simulated accurately, we find that accurately simulating either ET or SF along with SM is associated with significant trade-offs. Analyzing the trade-offs between the model responses (Table 7), we find that the higher trade-offs are mainly due to the fact that ET cannot be simulated accurately by calibrating the model with the other fluxes or storage variable. However, calibrating the model with ET produces lower error for SF (Fig. 5). This highlights the advantage of using a Pareto-based calibration approach, which does not assume any subjective weights in its objectives. Unlike deterministic calibration methodologies, we use parameter distributions from DREAM and AMALGAM to identify the parameters that cause significant trade-offs in accuracy between simulated fluxes and a storage variable. In addition to sensitive parameters that influence the behavioral simulation of model responses, we identify parameters that influence the trade-offs. For example, in the case of the Noah-MP model tested in this study, we see that the runoff parameter REFDK and the exponent in the Brooks-Corey equation (BB) influence the trade-off between ET and SM. This not only shows the advantages of framing multivariate calibration as a Pareto optimality problem but also highlights the fact that relatively insensitive parameters (such as REFDK for ET and SM) can exert a big influence on the accurate simulation of multiple water balance components.

We note that the results and conclusions we present in this study are for a specific combination of hydrologic model, input datasets, reference data and model parameters. For a different hydrologic model or calibration approach, the value of incorporating different water balance components and the relationship between them may be different. For example, the time period considered for calibration is one year and the study area is the Mississippi river basin. The calibrated parameters may not be applicable to a river basin with different physical characteristics and hydroclimatic conditions. The higher errors in the simulated water balance components (Fig. 4) may be due to the chosen calibration period of one year. This is especially true for streamflow, as only 72 data points (six basins and 12 months) were used for calibration as opposed to around 60,000 data points (all the grid cells) for ET and SM. Increasing the calibration period or the number of parameters may lead to improved accuracy in the simulation of the water balance components, and hence lower behavioral thresholds and improved Pareto optimal solutions. However, the multivariate calibration framework developed in this study is model- and data-independent and can be used to analyze the value of every flux or state variable under consideration. Future work involves extending the current study in several ways. First, we wish to incorporate more than two water balance components into the calibration to see if it changes the nature of trade-offs. For example, in the case of ET and SM, it would be interesting to see if incorporating SF reduces the overall trade-offs in accuracy among the water balance components. As REFDK is a runoff-related parameter that affects the trade-off between ET and SM, incorporating SF could lead to better discovery of REFDK. Second, incorporating estimates of total water storage (TWS) could yield better performance than the near-surface soil moisture estimates used in this study. Third, applying the developed framework to different hydrologic models and conducting inter-model comparison studies would help in model selection. Finally, the developed framework can help in the development, testing and diagnosis of new hydrologic models and model hypotheses from a multivariate perspective.

Acknowledgments

We acknowledge funding support from the NASA Applied Sciences Water Resources Application Grant # NNX15AC33G. We thank Jasper Vrugt for help with DREAM and AMALGAM. We also would like to thank two anonymous reviewers for their in-depth review and constructive comments. The GLEAM ET dataset has been sourced from www.gleam.eu, ESA-CCI soil moisture from www.esa-soilmoisture-cci.org, Ameriflux ET from https://ameriflux.lbl.gov/ and TAMU soil moisture from http://soilmoisture.tamu.edu/Data/. NASA's Land Information System is sourced from https://lis.gsfc.nasa.gov/. DREAM and AMALGAM software are available at http://faculty.sites.uci.edu/jasper/software/. HUC-2 basin runoff, model configuration files and post-process scripts are openly available at https://ucla.box.com/v/data-adwr-calibration.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.advwatres.2019.06.005.

Appendix A. Validation of GLEAM ET and ESA-CCI SM

Remotely sensed hydrologic fluxes and state variables such as precipitation, ET and SM are subject to large uncertainties due to differences in retrieval algorithms and sensors (Kidd and Huffman, 2011; Gebregiorgis and Hossain, 2014), thus necessitating the need for validating the chosen ET and SM datasets with ground-based measurements. We compare the GLEAM estimates with ground-based flux tower measurements from the Ameriflux network (https://ameriflux.lbl.gov/). A scatter plot of GLEAM versus Ameriflux ET for 2004 shows that GLEAM is capable of representing the ET flux in all the sub-basins to a fair degree of accuracy (Fig. A1a). The RMSE of GLEAM data is estimated to be 21.4 mm/month. For validating ESA-CCI SM data, we make use of near-surface soil moisture measurements from the TAMU North American Soil Moisture Database (NASMDB) (Quiring et al., 2016). For the study period, SM sensors from only three of the six sub-basins are available. The scatter plot (Fig. A1b) shows that remote sensing data overestimates the observed soil moisture in the Lower Mississippi Region. The RMSE value for ESA-CCI SM is 0.12 m^3/m^3.

As these satellite-based datasets are used together in multivariate calibration, we quantify the error in the closure of water balance (Fig. A1c

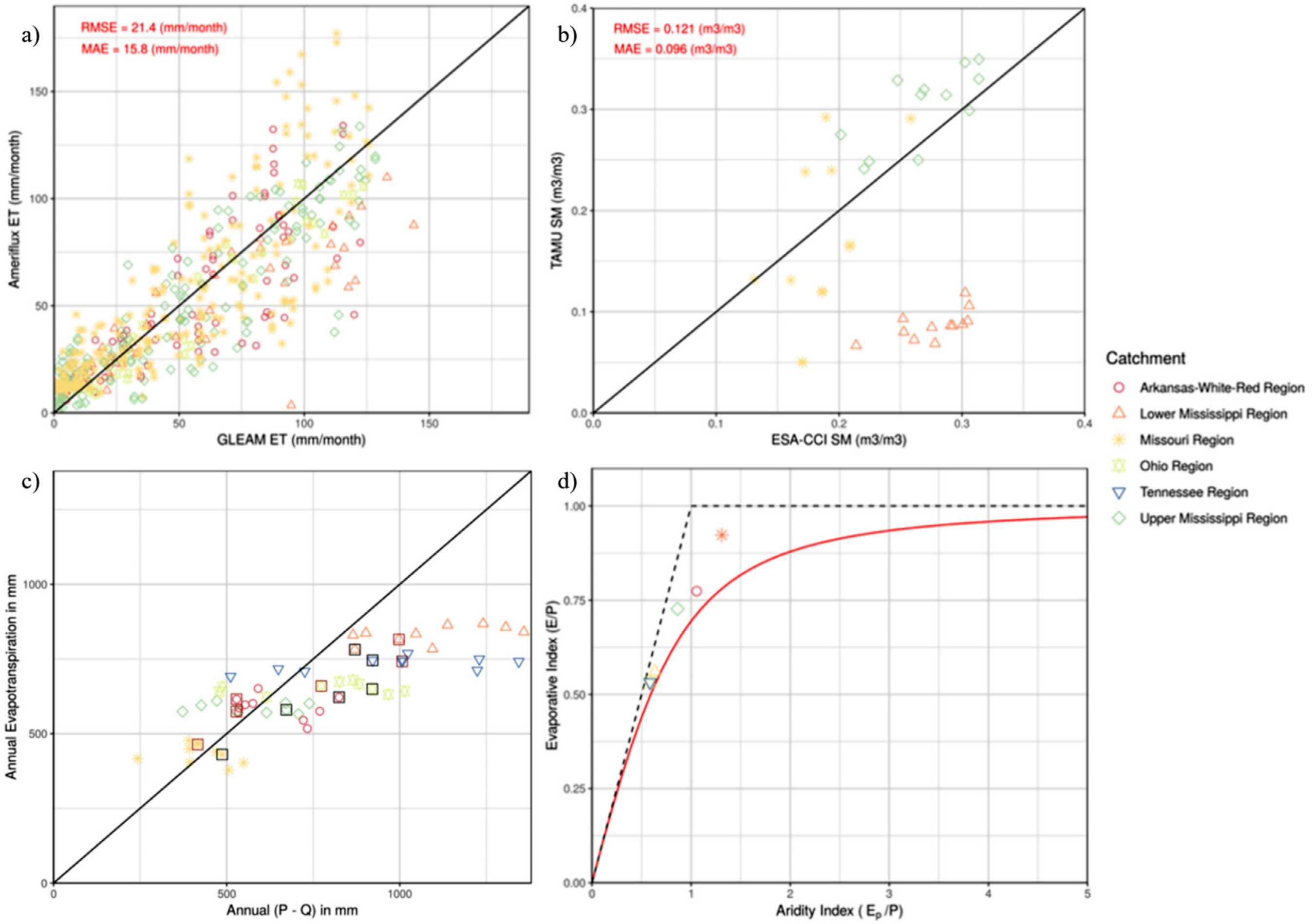

Fig. A1. Scatter plots of (a) GLEAM ET vs Ameriflux measurements (top right panel); (b) ESA-CCI soil moisture vs TAMU NASMDB measurements (bottom panel) for 2004; (c) annual GLEAM ET vs Annual Precipitation (P) – Runoff (Q) for the years 2000–2009 (bottom left). The errors for the calibration and validation years (2004 and 2005) are highlighted by black and brown bounding boxes respectively; and (d) the Budyko space (Evaporative index vs Aridity index) averaged over the years 2000–2009 for the six catchments. The red line is the ideal catchment water-energy balance as represented by the Budyko hypothesis. The dotted lines represent the water (horizontal line) and energy (diagonal line) limits (bottom right). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

and A1d). First, we compare the annual ET over the six HUC-2 catchments in the study area with the difference between precipitation (P) and runoff (Q) for the years 2000–2009. From Fig. A1c, it is evident that the errors in the closure of water balance are quite low for most of the catchments. The exceptions are three years in the Lower Mississippi and Tennessee regions in which inter-annual storage (soil moisture and groundwater) changes may play an important role. Importantly, the water balance closure errors for the calibration and validation periods are low across all the regions, including the Lower Mississippi and Tennessee regions. The mean annual water balance closure error from reference datasets (P - ET - Q), averaged over the entire Mississippi basin, is about 108 mm/year (9 mm/month). When the water balance components are summed over the entire basin, the water balance closure error is approximately 640 mm (53 mm/month). We also make sure that the reference datasets do not exceed catchment-scale water and energy limits as described by the Budyko hypothesis. In this study we make use of Fu's equation (Fu, 1981), a single parameter Budyko function that relates the evaporative index (E/P) and the aridity index (E_p /P, E_p is potential ET) as

$$\frac{E}{P} = 1 + \frac{E_p}{P} - \left(1 + \left(\frac{E_p}{P}\right)^{\omega}\right)^{\frac{1}{\omega}} \tag{6}$$

where ω is the Budyko parameter that has no analytic solution. We use a generally accepted representative value of 2.6 in Eq. (6) to construct the Budyko curve (red line in Fig. A1d). We see that all catchments are closely clustered around the Budyko curve except for the Missouri region. All catchments are also within the energy and water limits (dotted lines in Fig. A1d).

References

Ball, J.T., Woodrow, I.E., Berry, J.A., 1987. A model predicting stomatal conductance and its contribution to the control of photosynthesis under different environmental conditions. In: Progress in Photosynthesis Research. Springer, Netherlands, Dordrecht, pp. 221–224. https://doi.org/10.1007/978-94-017-0519-6_48.

Beven, K., 1996. Equifinality and uncertainty in geomorphological modelling. In: The Scientific Nature of Geomorphology: Proceedings of the Twenty-Seventh Binghamton Symposium in Geomorphology, Held 27-29 September, 1996, 27. John Wiley & Sons, p. 289.

Beven, K., 2001. How far can we go in distributed hydrological modelling? Hydrol. Earth Syst. Sci. 5 (1), 1–12. https://doi.org/10.5194/hess-5-1-2001.

Beven, K., 2006. A manifesto for the equifinality thesis. J. Hydrol. 320 (1), 18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007.

Beven, K., Binley, A., 1992. The future of distributed models: model calibration and uncertainty prediction. Hydrol. Processes 6 (3), 279–298. https://doi.org/10.1002/hyp.3360060305.

Beven, K.J., 2018. On hypothesis testing in hydrology: why falsification of models is still a really good idea. Wiley Interdiscip. Rev.: Water 5 (3), e1278. https://doi.org/10.1002/wat2.1278.

Budyko, M., 1974. *Climate and Life*, International Geophysics Series. Academic Press.

Cai, X., Yang, Z.-L., David, C.H., Niu, G.-Y., Rodell, M., 2014. Hydrological evaluation of the Noah-MP land surface model for the Mississippi river basin. J. Geophys. Res.: Atmos. 119 (1), 23–38. https://doi.org/10.1002/2013JD020792, 2013JD020792.

Chen, F., Janjić, Z., Mitchell, K., 1997. Impact of atmospheric surface-layer parameterizations in the new land-surface scheme of the NCEP mesoscale Eta model. Bound. Layer Meteorol. 85 (3), 391–421. https://doi.org/10.1023/A:1000531001463.

Christensen, N.S., Lettenmaier, D.P., 2007. A multimodel ensemble approach to assessment of climate change impacts on the hydrology and water resources of the Colorado river basin. Hydrol. Earth Syst. Sci. 11 (4), 1417–1434. https://doi.org/10.5194/hess-11-1417-2007.

Cosby, B.J., Hornberger, G.M., Clapp, R.B., Ginn, T.R., 1984. A statistical exploration of the relationships of soil moisture characteristics to the physical properties of soils. Water Resour. Res. 20 (6), 682–690. https://doi.org/10.1029/WR020i006p00682.

Cuntz, M., Mai, J., Samaniego, L., Clark, M., Wulfmeyer, V., Branch, O., Attinger, S., Thober, S., 2016. The impact of standard and hard-coded parameters on the hydrologic fluxes in the Noah-MP land surface model. J Geophys. Res.: Atmos. 121 (18). https://doi.org/10.1002/2016JD025097, 10,676–10,7002016JD025097.

Cuo, L., Zhang, Y., Gao, Y., Hao, Z., Cairang, L., 2013. The impacts of climate change and land cover/use transition on the hydrology in the Upper Yellow river basin, China. J. Hydrol. 502 (Supplement C), 37–52. https://doi.org/10.1016/j.jhydrol.2013.08.003.

Deb, K., Pratap, A., Agarwal, S., Meyarivan, T., 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans. Evol. Comput. 6 (2), 182–197. https://doi.org/10.1109/4235.996017.

Derber, J.C., Parrish, D.F., Lord, S.J., 1991. The new global operational analysis system at the national meteorological center. Weather Forecast. 6 (4), 538–547.

Dorigo, W., Wagner, W., Albergel, C., Albrecht, F., Balsamo, G., Brocca, L., Chung, D., Ertl, M., Forkel, M., Gruber, A., Haas, E., Hamer, P.D., Hirschi, M., Ikonen, J., de Jeu, R., Kidd, R., Lahoz, W., Liu, Y.Y., Miralles, D., Mistelbauer, T., Nicolai-Shaw, N., Parinussa, R., Pratola, C., Reimer, C., van der Schalie, R., Seneviratne, S.I., Smolander, T., Lecomte, P., 2017. ESA CCI soil moisture for improved earth system understanding: State-of-the art and future directions. Remote Sens. Environ. 203 (Supplement C), 185–215. https://doi.org/10.1016/j.rse.2017.07.001.

Efstratiadis, A., Koutsoyiannis, D., 2010. One decade of multi-objective calibration approaches in hydrological modelling: a review. Hydrol. Sci. J. 55 (1), 58–78. https://doi.org/10.1080/02626660903526292.

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

Fu, B., 1981. On the calculation of the evaporation from land surface. Sci. Atmos. Sin 5 (1), 23–31.

Gebregiorgis, A., Hossain, F., 2014. Making satellite precipitation data work for the developing world. IEEE Geosci. Remote Sens. Mag. 2 (2), 24–36. https://doi.org/10.1109/MGRS.2014.2317561.

Gupta, H.V., Bastidas, L.A., Sorooshian, S., Shuttleworth, W.J., Yang, Z.L., 1999. Parameter estimation of a land surface scheme using multicriteria methods. J Geophys. Res.: Atmos. 104 (D16). https://doi.org/10.1029/1999JD900154, 19,491–19,503.

Gupta, H.V., Sorooshian, S., Yapo, P.O., 1998. Toward improved calibration of hydrologic models: multiple and noncommensurable measures of information. Water Resour. Res. 34 (4), 751–763. https://doi.org/10.1029/97WR03495.

Gupta, H.V., Wagener, T., Liu, Y., 2008. Reconciling theory with observations: elements of a diagnostic approach to model evaluation. Hydrol. Processes 22 (18), 3802–3813. https://doi.org/10.1002/hyp.6989.

Haario, H., Saksman, E., Tamminen, J., 2001. An adaptive metropolis algorithm. Bernoulli 7 (2), 223–242. https://doi.org/10.2307/3318737.

Hogue, T.S., Bastidas, L.A., Gupta, H.V., Sorooshian, S., 2006. Evaluating model performance and parameter behavior for varying levels of land surface model complexity. Water Resour. Res. 42 (8). https://doi.org/10.1029/2005WR004440.

Immerzeel, W., Droogers, P., 2008. Calibration of a distributed hydrological model based on satellite evapotranspiration. J. Hydrol. 349 (3), 411–424. https://doi.org/10.1016/j.jhydrol.2007.11.017.

Jordan, R., 1991. A One-Dimensional Temperature Model for a Snow Cover: Technical Documentation for SNTHERM.89. Cold Regions Research and Engineering Lab, Hanover NH Tech. Rep..

Kavetski, D., Kuczera, G., Franks, S.W., 2006. Bayesian analysis of input uncertainty in hydrological modeling: 1. theory. Water Resour. Res. 42 (3). https://doi.org/10.1029/2005WR004368.

Kennedy, J., Eberhart, R.C., 2001. Swarm Intelligence. Morgan Kaufmann Publishers Inc, San Francisco, CA, USA.

Khu, S.T., Madsen, H., 2005. Multiobjective calibration with Pareto preference ordering: an application to rainfall-runoff model calibration. Water Resour. Res. 41 (3). https://doi.org/10.1029/2004WR003041.

Kidd, C., Huffman, G., 2011. Global precipitation measurement. Meteorol. Appl. 18 (3), 334–353. https://doi.org/10.1002/met.284.

Koppa, A., Gebremichael, M., 2017. A framework for validation of remotely sensed precipitation and evapotranspiration based on the Budyko hypothesis. Water Resour. Res. 53 (10), 8487–8499. https://doi.org/10.1002/2017WR020593.

Kumar, S., Peters-Lidard, C., Tian, Y., Houser, P., Geiger, J., Olden, S., Lighty, L., Eastman, J., Doty, B., Dirmeyer, P., Adams, J., Mitchell, K., Wood, E., Sheffield, J., 2006. Land information system: an interoperable framework for high resolution land surface modeling. Environ. Model. Softw. 21 (10), 1402–1415. https://doi.org/10.1016/j.envsoft.2005.07.004.

Laloy, E., Rogiers, B., Vrugt, J.A., Mallants, D., Jacques, D., 2013. Efficient posterior exploration of a high-dimensional groundwater model from two-stage Markov chain Monte Carlo simulation and polynomial chaos expansion. Water Resour. Res. 49 (5), 2664–2682. https://doi.org/10.1002/wrcr.20226.

Leng, G., Tang, Q., Rayburg, S., 2015. Climate change impacts on meteorological, agricultural and hydrological droughts in China. Global Planetary Change 126 (Supplement C), 23–34. https://doi.org/10.1016/j.gloplacha.2015.01.003.

Lettenmaier, D.P., Alsdorf, D., Dozier, J., Huffman, G.J., Pan, M., Wood, E.F., 2015. Inroads of remote sensing into hydrologic science during the WRR era. Water Resour. Res. 51 (9), 7309–7342. https://doi.org/10.1002/2015WR017616.

Li, D., Wrzesien, M.L., Durand, M., Adam, J., Lettenmaier, D.P., 2017. How much runoff originates as snow in the western United States, and how will that change in the future? Geophys. Res. Lett. 44 (12), 6163–6172. https://doi.org/10.1002/2017GL073551, 2017GL073551.

López López, P., Sutanudjaja, E.H., Schellekens, J., Sterk, G., Bierkens, M.F.P., 2017. Calibration of a large-scale hydrological model using satellite-based soil moisture and evapotranspiration products. Hydrol. Earth Syst. Sci. 21 (6), 3125–3144. https://doi.org/10.5194/hess-21-3125-2017.

Ma, N., Niu, G.-Y., Xia, Y., Cai, X., Zhang, Y., Ma, Y., Fang, Y., 2017. A systematic evaluation of Noah-MP in simulating land-atmosphere energy, water, and carbon exchanges over the continental United States. J. Geophys. Res.: Atmos. 122 (22). https://doi.org/10.1002/2017JD027597, 12,245–12,268.

Madsen, H., 2003. Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives. Adv. Water Resour. 26 (2), 205–216. https://doi.org/10.1016/S0309-1708(02)00092-1.

Martens, B., Miralles, D.G., Lievens, H., van der Schalie, R., de Jeu, R.A.M., Fernández-Prieto, D., Beck, H.E., Dorigo, W.A., Verhoest, N.E.C., 2017. Gleam v3: satellite-based land evaporation and rootzone soil moisture. Geosci. Model. Dev. 10 (5), 1903–1925. https://doi.org/10.5194/gmd-10-1903-2017.

Mendoza, P.A., Clark, M.P., Mizukami, N., Newman, A.J., Barlage, M., Gutmann, E.D., Rasmussen, R.M., Rajagopalan, B., Brekke, L.D., Arnold, J.R., 2015. Effects of hydrologic model choice and calibration on the portrayal of climate change impacts. J. Hydrometeorol. 16, 762–780. https://doi.org/10.1175/JHM-D-14-0104.1.

Middelkoop, H., Daamen, K., Gellens, D., Grabs, W., Kwadijk, J.C.J., Lang, H., Parmet, B.W.A.H., Schädler, B., Schulla, J., Wilke, K., 2001. Impact of climate change on hydrological regimes and water resources management in the Rhine basin. Climatic Change 49 (1), 105–128. https://doi.org/10.1023/A:1010784727448.

Niu, G.-Y., Yang, Z.-L., 2006. Effects of frozen soil on snowmelt runoff and soil water storage at a continental scale. J. Hydrometeorol. 7 (5), 937–952. https://doi.org/10.1175/JHM538.1.

Niu, G.-Y., Yang, Z.-L., Dickinson, R.E., Gulden, L.E., Su, H., 2007. Development of a simple groundwater model for use in climate models and evaluation with gravity recovery and climate experiment data. J. Geophys. Res.: Atmos. 112 (D7). https://doi.org/10.1029/2006JD007522.

Niu, G.-Y., Yang, Z.-L., Mitchell, K.E., Chen, F., Ek, M.B., Barlage, M., Kumar, A., Manning, K., Niyogi, D., Rosero, E., Tewari, M., Xia, Y., 2011. The community Noah land surface model with multiparameterization options (Noah-MP): 1. model description and evaluation with local-scale measurements. J. Geophys. Res.: Atmos. 116 (D12). https://doi.org/10.1029/2010JD015139, n/a–n/ad12109.

Quiring, S.M., Ford, T.W., Wang, J.K., Khong, A., Harris, E., Lindgren, T., Goldberg, D.W., Li, Z., 2016. The North American soil moisture database: development and applications. Bull. Am. Meteorol. Soc. 97 (8), 1441–1459. https://doi.org/10.1175/BAMS-D-13-00263.1.

Rakovec, O., Kumar, R., Mai, J., Cuntz, M., Thober, S., Zink, M., Attinger, S., Schäfer, D., Schrön, M., Samaniego, L., 2016a. Multiscale and multivariate evaluation of water fluxes and states over European river basins. J. Hydrometeorol. 17 (1), 287–307. https://doi.org/10.1175/JHM-D-15-0054.1.

Rakovec, O., Kumar, R., Attinger, S., Samaniego, L., 2016b. Improving the realism of hydrologic model functioning through multivariate parameter estimation. Water Resour. Res. 52 (10), 7779–7792. https://doi.org/10.1002/2016WR019430.

Rientjes, T., Muthuwatta, L., Bos, M., Booij, M., Bhatti, H., 2013. Multi-variable calibration of a semi-distributed hydrological model using streamflow data and satellite-based evapotranspiration. J. Hydrol. 505, 276–290. https://doi.org/10.1016/j.jhydrol.2013.10.006.

Sadegh, M., Vrugt, J.A., 2014. Approximate Bayesian computation using Markov chain Monte Carlo simulation: DREAM(abc). Water Resour. Res. 50 (8), 6767–6787. https://doi.org/10.1002/2014WR015386.

Samaniego, L., Kumar, R., Attinger, S., 2010. Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. Water Resour. Res. 46 (5). https://doi.org/10.1029/2008WR007327.

Samaniego, L., Kumar, R., Thober, S., Rakovec, O., Zink, M., Wanders, N., Eisner, S., Müller Schmied, H., Sutanudjaja, E.H., Warrach-Sagi, K., Attinger, S., 2017. Toward seamless hydrologic predictions across spatial scales. Hydrol. Earth Syst. Sci. 21 (9), 4323–4346. https://doi.org/10.5194/hess-21-4323-2017.

Schoups, G., Vrugt, J.A., 2010. A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. Water Resour. Res. 46 (10). https://doi.org/10.1029/2009WR008933.

Shafii, M., Tolson, B., Matott, L.S., 2014. Uncertainty-based multi-criteria calibration of rainfall-runoff models: a comparative study. Stochastic Environ. Res. Risk Assess. 28 (6), 1493–1510. https://doi.org/10.1007/s00477-014-0855-x.

Sheffield, J., Goteti, G., Wen, F., Wood, E.F., 2004. A simulated soil moisture based drought analysis for the united states. J. Geophys. Res.: Atmos. 109 (D24). https://doi.org/10.1029/2004JD005182, n/a–n/ad24108.

Storn, R., Price, K., 1997. Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. J. Global Optim. 11 (4), 341–359. https://doi.org/10.1023/A:1008202821328.

Sutanudjaja, E.H., van Beek, L.P.H., de Jong, S.M., van Geer, F.C., Bierkens, M.F.P., 2013. Calibrating a large-extent high-resolution coupled groundwater-land surface model using soil moisture and discharge data. Water Resour. Res. 50 (1), 687–705. https://doi.org/10.1002/2013WR013807.

Verseghy, D.L., McFarlane, N.A., Lazare, M., 1991. CLASS – a Canadian land surface scheme for GCMs, II. Vegetation model and coupled runs. Int. J. Climatol. 13 (4), 347–370. https://doi.org/10.1002/joc.3370130402.

Vrugt, J.A., 2016. Markov chain Monte Carlo simulation using the DREAM software package: theory, concepts, and MATLAB implementation. Environ. Model. Softw. 75, 273–316. https://doi.org/10.1016/j.envsoft.2015.08.013.

Vrugt, J.A., Robinson, B.A., 2007. Improved evolutionary optimization from genetically adaptive multimethod search. Proc. Natl. Acad. Sci. USA 104 (3), 708–711. https://doi.org/10.1073/pnas.0610471104.

Vrugt, J.A., Sadegh, M., 2013. Toward diagnostic model calibration and evaluation: approximate Bayesian computation. Water Resour. Res. 49 (7), 4335–4345. https://doi.org/10.1002/wrcr.20354.

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

Vrugt, J.A., ter Braak, C.J.F., Gupta, H.V., Robinson, B.A., 2009a. Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling? Stochastic Environ. Res. Risk Assess. 23 (7), 1011–1026. https://doi.org/10.1007/s00477-008-0274-y.

Vrugt, J.A., Ter Braak, C., Diks, C., Robinson, B.A., Hyman, J.M., Higdon, D., 2009b. Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. Int. J. Nonlinear Sci. Numer. Simul. 10 (3), 273–290.

Wanders, N., Bierkens, M.F.P., de Jong, S.M., de Roo, A., Karssenberg, D., 2014. The benefits of using remotely sensed soil moisture in parameter identification of large-scale hydrological models. Water Resour. Res. 50 (8), 6874–6891. https://doi.org/10.1002/2013WR014639.

Yang, R., Friedl, M.A., 2003. Modeling the effects of three-dimensional vegetation structure on surface radiation and energy balance in boreal forests. J. Geophys. Res.: Atmos. 108 (D16). https://doi.org/10.1029/2002JD003109.

Zink, M., Mai, J., Cuntz, M., Samaniego, L., 2018. Conditioning a hydrologic model using patterns of remotely sensed land surface temperature. Water Resour. Res. 54. https://doi.org/10.1002/2017WR021346.