Hydrol. Earth Syst. Sci., 16, 3029–3048, 2012
www.hydrol-earth-syst-sci.net/16/3029/2012/
doi:10.5194/hess-16-3029-2012

Hydrology and Earth System Sciences

Multi-criteria parameter estimation for the Unified Land Model

B. Livneh and D. P. Lettenmaier

University of Washington Department of Civil and Environmental Engineering, P.O. Box 352700, Seattle, WA 98195, USA

Correspondence to: B. Livneh (ben.livneh@gmail.com)

Received: 12 March 2012 – Published in Hydrol. Earth Syst. Sci. Discuss.: 3 April 2012
Revised: 31 July 2012 – Accepted: 1 August 2012 – Published: 29 August 2012

Abstract. We describe a parameter estimation framework for the Unified Land Model (ULM) that utilizes multiple independent data sets over the continental United States. These include a satellite-based evapotranspiration (ET) product based on MODerate resolution Imaging Spectroradiometer (MODIS) and Geostationary Operational Environmental Satellites (GOES) imagery, an atmospheric-water balance based ET estimate that utilizes North American Regional Reanalysis (NARR) atmospheric fields, terrestrial water storage content (TWSC) data from the Gravity Recovery and Climate Experiment (GRACE), and streamflow (Q) primarily from the United States Geological Survey (USGS) stream gauges. The study domain includes 10 large-scale ($\geq 10^5$ km^2) river basins and 250 smaller-scale ($< 10^4$ km^2) tributary basins. ULM, which is essentially a merger of the Noah Land Surface Model and Sacramento Soil Moisture Accounting Model, is the basis for these experiments. Calibrations were made using each of the data sets individually, in addition to combinations of multiple criteria, with multi-criteria skill scores computed for all cases. At large scales, calibration to Q resulted in the best overall performance, whereas certain combinations of ET and TWSC calibrations lead to large errors in other criteria. At small scales, about one-third of the basins had their highest Q performance from multi-criteria calibrations (to Q and ET) suggesting that traditional calibration to Q may benefit by supplementing observed Q with remote sensing estimates of ET. Model streamflow errors using optimized parameters were mostly due to over (under) estimation of low (high) flows. Overall, uncertainties in remote-sensing data proved to be a limiting factor in the utility of multi-criteria parameter estimation.

1 Introduction

The evolution of land surface models (LSMs) towards increasingly complex representations of hydrologic and biophysical processes requires special attention to the fidelity of the models in partitioning water and energy budget components. The traditional validation of models using observations of a single prognostic variable can result in model predictions that are inherently biased towards that variable (McCabe et al., 2005). The evaluation of multiple model outputs (as opposed to single-output analysis, such as streamflow) has received increasing attention (e.g. Gupta et al., 1999; Crow et al., 2003; McCabe et al., 2005; Khu et al., 2008; Werth and Güntner, 2010; Milzow et al., 2011). Among the variables other than streamflow that have been used for LSM evaluation are evapotranspiration (Nandagiri, 2007), surface heat fluxes (Gupta et al., 1999; McCabe et al., 2005), hydrochemical and isotope tracers (Son and Sivapalan, 2007; Lischeid, 2008; Birkel et al., 2010), land surface temperature (Crow et al., 2003; McCabe et al., 2005), remotely sensed soil moisture (Brocca et al., 2010; Milzow et al., 2011), snow water equivalent (MacLean et al., 2010), terrestrial water storage (Werth and Güntner, 2010; Milzow et al., 2010), and water table level (Khu et al., 2008). The more frequent use of multivariate observations is attributable in part to their growing availability. Some satellite-based observations now have periods of record exceeding a decade for single sensors, and multiple decades for some multi-sensor merged records.

In the context of parameter estimation, multi-criteria analyses can aid in addressing the issue of equifinality (Beven and Freer, 2001). The equifinality problem arises when different parameter sets result in similar model performance. One approach to reducing equifinality issues and quantifying uncertainties in model calibration is the Generalized Likelihood Uncertainty Estimation (GLUE) framework of Beven

Published by Copernicus Publications on behalf of the European Geosciences Union.

and Binley (1992), which can aid in selection of model calibration parameters through estimating the likelihood that each parameter set is the true predictor of the system. A distribution of likelihoods among many parameter values is then generated and used to define uncertainties and select parameters. Alternatively, the degree of system complexity resolvable by a model has been proposed as a diagnostic tool (Yilmaz et al., 2008; Gupta et al., 2008) that could be applied in a Bayesian uncertainty context or for assessing model structural and behavioral consistency.

Herein we consider an alternative calibration methodology (detailed in Sect. 3.4) that compares model performance when parameters are selected via combinations of ancillary objective functions. The addition of observational sources is used to constrain parameter values – i.e. multivariate performance analyses – such that parameter combinations that produce unrealistic results for certain combinations of simultaneous criteria can be discarded. Despite the multivariate performance analysis, we selected a single best parameter set in order to compare performance across calibrations using differing combinations of criteria (described in Sect. 3.4). Robust model parameters are especially important when models are used to predict outcomes for model forcings outside the range observed in the model parameter estimation (calibration) period. The interannual variability of streamflow regimes is one such example, which provides a basis for the investigation of potential future changes in river discharge that might result from climate or land-use change (Kingston et al., 2011). Robust model parameters are also essential for examining the importance of spatial and temporal scale on land surface response. Spatial scale can, for instance, determine the nature of environmental impact assessments (João, 2002), and the categorization of droughts (Shukla et al., 2011), but also determine how localized hydrologic events propagate through a larger system (for instance, flash flooding from tributary catchments as it affects the hydrologic response of a much larger region). It is therefore anticipated that assessing model performance against independent observational data at a range of spatial scales will be insightful for selecting representative parameters.

2 Modeling context

The Unified Land Model (ULM – Livneh et al., 2011) is the LSM used in this study. ULM is essentially a merger of two widely used models: the Noah LSM (Ek et al., 2003; used in most of the National Oceanic and Atmospheric Administration (NOAA) coupled weather and climate models) and the Sacramento Soil Moisture Accounting Model (Sac; Burnash et al., 1973; used for hydrologic prediction within the National Weather Service). The parameter estimation experiments reported here can also be viewed as a means to evaluate ULM rigorously in ways that extend the work of Livneh et al. (2011). Additionally, given the ULM's heritage and widespread use of Noah and Sac, the implications of the results should be broadly relevant to the modeling community.

The objective of this work is to examine the benefits and potential tradeoffs of incorporating multiple observations (multiple-criteria) into model calibration across a range of hydroclimatic conditions and spatial scales. This will involve computing simultaneous skill-scores between the model and each observed criterion. Using this information, the nature of error accumulations and interannual variability in resulting model predictions can also be examined.

We first apply a multivariate model calibration procedure over some of the major river basins of the continental United States (CONUS), and follow with similar calibrations for selected interior tributary catchments. Single and multi-criteria objective functions were used to assess the added value of including information such as remotely sensed ET and TWSC in the calibration procedure. These criteria were selected to provide at least one observation-based data source for each component of the water budget. Estimated parameters were then used to analyze simulated streamflow variability, seasonality, and autocorrelation, examining both model skill and error propagation across different spatial scales and hydroclimatic regions.

3 Data and methods

In this section we describe the experimental design, including the study domain, the model, and model forcing and evaluation data. We follow with a description of the model calibration strategy and the trend and error analyses.

3.1 Basin selection, streamflow, and meteorological data

The study domain is comprised of river basins of different sizes within the CONUS, selected to provide a broad cross section of hydroclimatic conditions and basin areas that are representative of typical land surface modeling applications. The largest river basins (hereafter major basins) are shown in Fig. 1, and their characteristics are summarized in Table 1. For several major basins, particularly in the western US, naturalized streamflow data were obtained that have been adjusted for anthropogenic impacts, including upstream (reservoir) regulation, water withdrawals and evaporation from upstream reservoirs (see Table 1). In addition to the 10 major basins, a set of 250 smaller catchments (herein tributaries) was selected, most of which are tributaries to the major basins (Fig. 2). The tributaries are a subset of the model parameter estimation experiment (MOPEX; Schaake et al., 2006) data set, which have been screened to assure that they have an adequate density of precipitation gauges and are minimally affected by upstream anthropogenic activities such as irrigation diversion and reservoir operations. An alternative screening tool (not used here) is the more recent GAGES-II database (Falcone et al., 2010), which also

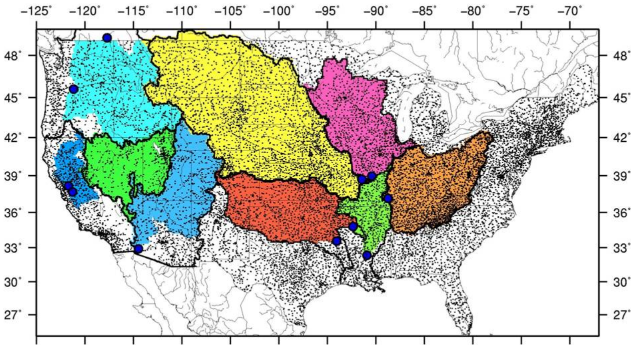

Fig. 1. Large-scale study domain, including precipitation gauges (black dots), as well as major hydrologic regions (shaded) that are defined through their drainage at stream gauges (blue circles). The un-shaded areas within these regions are either downstream of the stream gauge, or consist of many smaller river basins which drain directly into the Atlantic or Pacific Oceans or the Gulf of Mexico.

classifies basins based on anthropogenic disturbance. Hence, streamflow observations for the tributaries were obtained directly from United States Geologic Survey (USGS) archives. All basins were further screened here to have a minimum of 20 yr of data with 100 % record completeness within the period 1990–2009 to facilitate the use of remote sensing data sets in multi-criteria parameter estimation.

The meteorological data used in this study were derived by Livneh et al. (2012) and are available at a 0.0625° resolution over the CONUS domain for the period 1915–2010. Precipitation and daily minimum and maximum temperatures were obtained for the NOAA Cooperative Observer (Co-op) stations shown in Fig. 1. Precipitation and temperature were gridded directly from station data. Wind data were linearly interpolated from a larger (1.9° latitude-longitude) NCEP–NCAR reanalysis grid (Kalnay et al., 1996) that was used to produce daily wind climatology for years prior to 1948. For complete details of model forcing data, see Livneh et al. (2012).

3.2 Auxiliary model evaluation data

In addition to streamflow observations, we made use of two independent estimates of ET, which, like streamflow, are predicted by ULM. The first arises from an atmospheric water balance over the major basins, whereas the second, derived from remote sensing, is available on a spatially distributed basis, but for a relatively short (compared with most of the streamflow records) period of roughly one decade.

3.2.1 Atmospheric water balance ET ($\mathrm{ET_{AWB}}$)

Computing an atmospheric water balance has been a longstanding means for studying atmospheric exchanges of moisture over large areas. For a given atmospheric domain, with vertical extent to the 100 millibar height,

$$\nabla \cdot \frac{1}{g}\int\limits_{100}^{p_\mathrm{s}} \overline{qV}\,\mathrm{d}p + \frac{\partial}{\partial t}\left(\frac{1}{g}\int\limits_{100}^{p_\mathrm{s}} \overline{q}\,\mathrm{d}p\right) = P - \mathrm{ET} \qquad (1)$$

where the first term is the convergence of liquid into, or out of the column, the second term is the change in moisture (or precipitable water) from the column over time, q is the specific humidity, V is the mean horizontal wind velocity, p is pressure at elevation, g is the gravitational constant, and P is precipitation. Historically, the terms on the left-hand side of Eq. (1) were obtained using a "picket fence" approach based on radiosonde observations (e.g. Starr et al., 1965; Rasmussen, 1967; Rosen and Omolayo, 1981; Ropelewski and Yarosh, 1998). Areal moisture fluxes could then be estimated by integrating the divergence spatially over the domain, following Green's theorem. More recent studies (Oki et al., 1995; Yeh et al., 1998; Syed et al., 2005; Yeh and Famiglietti, 2008) have used this approach, where the spatial fields come from atmospheric reanalyses, which assimilate radiosonde data, as well as other satellite sources of information about the vertical profile of moisture and temperature. Yeh et al. (1998) examined the lower limit of spatial scale for applicability of the atmospheric water balance approach and found that, despite early estimates requiring areas $> 2 \times 10^6\,\mathrm{km}^2$ (Rasmussen, 1968), accurate estimation of the climatology of regional evaporation is possible at scales as small as $10^5\,\mathrm{km}^2$. At spatial scales smaller than about $10^5\,\mathrm{km}^2$, the accuracy of the estimates degrades rapidly.

We use the North American Regional Reanalysis (NARR; Mesinger et al., 2006) as the source of the two terms on the left-hand side of Eq. (1), both of which are standard NARR archived fields. The NARR output reflects the assimilation of radiosonde and satellite data that are routinely used in numerical weather prediction, but performed with a "frozen" version of the weather prediction model and data assimilation systems. The right-hand side of Eq. (1) is based on the gridded precipitation fields derived from a network of approximately 20 000 precipitation gauges across the CONUS by Livneh et al. (2012). Figure 3 illustrates the atmospheric water balance as used in this study.

3.3 Satellite-based ET ($\mathrm{ET_{SAT}}$)

Satellite remote sensing provides a promising alternative to direct observations for hydrologic prediction, although this is a source that has not been widely used to date – most likely because satellite-based data record lengths are only now approaching a decade. We used a MODIS-based ET data product produced by Tang et al. (2009). This product is based on the VI-Ts method described by Nishida et al. (2003) which uses only satellite-based (no surface data) products. Specifically, downward solar radiation is from the SRB data set of Pinker and Laszlo (1992), based on Geostationary Operational Environmental Satellites (GOES) and vegetation index (VI) and surface temperature (Ts) data are from MODIS.

Table 1. Major hydrologic regions considered in this study including streamflow gauges and drainage areas.

Hydrologic region	Abbreviation	Applicable criteria	Streamflow gauge location	USGS ID	Area (km^2)
Arkansas-Red	ARK	Q^1, ET, TWSC	Arkansas R. near Little Rock, AR	07263450	409 296
	RED	Q^1, ET, TWSC	Red R. at Index, AR	07337000	124 397
California	CALI	Q^2, ET, TWSC	Sacramento R. near Rio Vista, CA	11455420	69 300
			San Joaquin R. near Vernalis, CA	11303500	35 058
			Eastside streams and central valley floor	*	4655
Colorado	COLO	Q^3, ET, TWSC	Colorado R. above Imperial Dam, AZ	09429490	488 213
Columbia	CRB	Q^4, ET, TWSC	Columbia R. at Dalles, OR	14105700	613 827
			Columbia R. at Birchbank, BC	12323000	88 101
Great Basin	GBAS	ET, TWSC	N/A	N/A	367 602
Lower Mississippi	LOW	ET, TWSC	N/A	N/A	221 966
Upper Mississippi	UP	Q, ET, TWSC	Upper Mississippi R. at Grafton, IL	05587450	443 665
Missouri	MO	Q^5, ET, TWSC	Missouri R. at Hermann, MO	06934500	1 353 269
Ohio	OHIO	Q, ET, TWSC	Ohio R. at Metropolis, IL	03611500	525 768

[1] Naturalized streamflow data from the US Army Corps of Engineers,Tulsa OK office. [2] Naturalized streamflow data from California Data Exchange Commission. [3] Naturalized streamflow data from US Bureau of Reclamation, Lower Colorado Region. [4] Naturalized streamflow data from Columbia River Basin Climate Change Scenarios Database. [5] Naturalized streamflow data from US Army Corps of Engineers, Omaha NB office. * Unimpaired flow data for the Sacramento-San Joaquin River Delta were estimated by the California Department of Water Resources, which receives a small contribution from eastside streams and flows from the central valley floor. N/A – indicates that stream flow was not applicable; GBAS does not have an outlet at the basin boundary, LOW represents the confluence of multiple inflows and reliable flow data was not obtainable.

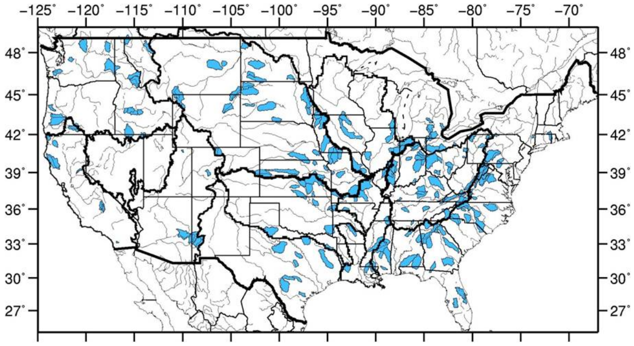

Fig. 2. Small-scale study domain comprised of 250 tributary catchments using USGS stream gauges that were screened to be minimally affected by diversions, with at least 20 yr of data in the past 3 decades to facilitate multi-criteria comparisons.

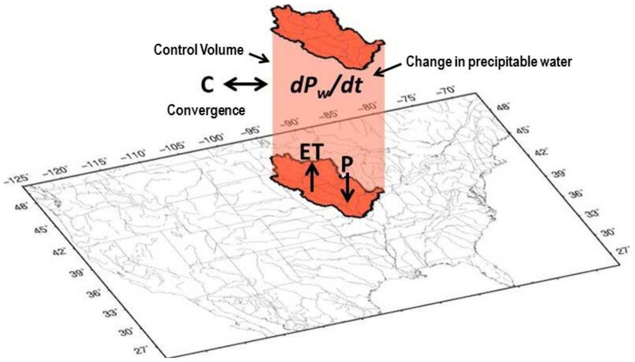

Fig. 3. Example schematic of the Upper Mississippi River Basin components needed to perform an atmospheric water balance to estimate ET (Eq. 1), including atmospheric moisture convergence, C, change in precipitable water, dP_w/dt, and precipitation, P.

Two key assumptions of the algorithm are (a) that the evaporative fraction is constant over the diurnal cycle, and is well estimated by values from the daytime satellite overpass (of EOS/Terra in this case), and (b) there is a substantial variation in VI-Ts pairs over a local region, such that an upper envelope of VI and Ts can be defined. The reader is referred to Tang et al. (2009) and Nishida et al. (2003) for details of the algorithm. The algorithm was applied at 0.05° spatial resolution, where each pixel represents the average of an area with 0.25° radius, to address assumption (b). To facilitate comparison with other criteria, a spatial average value of $\mathrm{ET_{SAT}}$ was computed for each basin.

Comparing this approach with ground observations, Tang et al. (2009) computed instantaneous and daily mean ET differences of less than 10 % and 15 % on average, respectively. VI-T_s derived ET agreed favorably with estimates from a much higher resolution Landsat-based method over irrigated areas of the Klamath River Basin in the western US. Nishida et al. (2003) found correlations of $R^2 > 0.85$ at 13 flux tower sites over CONUS. Kalma et al. (2008) surveyed a number of satellite-based ET methods (including the Nishida et al., 2003 VI-Ts method) and noted they can provide good estimates of the catchment's average evaporation on a daily basis subjected to cloud cover. However, they found that an important uncertainty in the ET estimates resulted from land surface temperature errors from the satellite estimates that could be as great as 3–5 K due to atmospheric effects. Ferguson et al. (2010) analyzed a similar satellite-based ET product and argued that a significant issue with satellite-based ET products is that they are not constrained by soil/surface water availability. They found that in some cases the high ET-demand during the warm season results in satellite-based ET estimates that are unrealistically large.

3.4 Terrestrial water storage change (TWSC)

The terrestrial water balance can be written as the difference between precipitation, P, and streamflow, Q, and ET:

$$\mathrm{TWSC} = P - Q - \mathrm{ET}. \tag{2}$$

Storage plays a key role in the Earth's climate system and the supply of freshwater for human use, via interaction with groundwater, soil moisture, plant water, snow, and land-ice. The Gravity Recovery and Climate Experiment (GRACE) provides a basis for estimating monthly variations of TWSC over areas on the order of $10^5\,\mathrm{km}^2$ based on the effect of TWSC on changes in the Earth's gravitational field measured by a pair of satellites. Temporal gravity variations at these spatial and temporal scales are mainly caused by mass redistribution in the atmosphere and oceans, tides, post-glacial rebound, and terrestrial water cycling (Klees et al., 2008). Monthly gravity field solutions are computed at the University of Texas at Austin Center for Space Research, the GeoForschungsZentrum Potsdam, and the Jet Propulsion Laboratory, which use different processing strategies and hence yield slightly different results. Similar to Werth and Güntner (2010), we used an average of GRACE gravity fields from these three processing centers (differences among the data sets can be considered a measure of data uncertainty), which were then averaged over each basin. Model-based TWSC was the sum of ULM-simulated total column soil moisture, and snow water equivalent (SWE) was compared with the GRACE product. Lo et al. (2010), Werth and Güntner (2010), and Milzow et al. (2011) have shown the potential for using GRACE-derived TWSC data in the calibration of LSMs. However, the GRACE record length is relatively short (from 2002), and the coarse spatial resolution complicates comparisons with model predictions for other than very large river basins.

3.5 Land surface model

Livneh et al. (2011) provide a complete description of ULM as used in this study. In general, the land surface components are from the Noah LSM – e.g. vegetation, ET computation, snow model, and algorithms for computing frozen soil, surface heat and radiative fluxes – whereas the subsurface elements (soil moisture and runoff generation algorithms, as well as infiltration) are from Sac. The snow model is described by Livneh et al. (2010). It essentially is the standard Noah snow model augmented to include time-varying albedo, partial snow cover, and retention of liquid water within the snowpack. Livneh et al. (2011) tested ULM at a small number of catchments and evaluated performance with respect to observed river discharge, flux towers measurements of surface heat fluxes, and soil moisture. Table 2 summarizes plausible physical ranges of the model soil parameters that constrained the parameter estimation here.

3.6 Calibration procedure and error analysis

By far the most common method for hydrologic model calibration is through minimization of differences between modeled and observed streamflow. The goal here was to extend this approach to include auxiliary observational data sources to evaluate and constrain model performance within a multi-criteria framework. The Nash-Sutcliffe efficiency (NSE – Nash and Sutcliffe, 1970) was chosen to quantify model performance. NSE is given as

$$\mathrm{NSE} = 1\frac{\sum_{t=1}^{n}(x_{\mathrm{s,t}} - x_{\mathrm{o,t}})^2}{\sum_{t=1}^{n}(x_{\mathrm{o,t}} - \mu_{\mathrm{o}})^2} = 1 - \frac{\mathrm{MSE}}{\sigma_{\mathrm{o}^2}} \tag{3}$$

where $x_{\mathrm{o,t}}$ and $x_{\mathrm{s,t}}$ are the observed and simulated values at each time step, μ_{o} is the observed mean and n is the total number of time steps. NSE is useful in comparing inter-basin performance, since it normalizes the mean squared error, MSE, by the observed variance, σ_{o}^2, of each basin, where an NSE value of 1 corresponds to a perfect model, while any value less than 0 describes a model that performs worse than simply using μ_{o} as the predictor. As described by Gupta et al. (2009), the NSE may be decomposed to represent the correlation between model and observed calibration variables (e.g. streamflow), difference of means, and difference of standard deviations between simulations and observations. They argue that calibrating a model within a multi-objective perspective towards these three components is preferred as it enables better hydrological interpretation of the solutions. Schaefli and Gupta (2007) discuss the difficulty in using NSE as a metric since it depends on the seasonality of the reference signal. Although we have not considered this in our implementation, our methodology could be easily modified to incorporate metrics that are less affected by seasonal variations.

We performed optimizations using the MOCOM-UA algorithm, first developed by Yapo et al. (1998), as a means of maximizing NSE (minimizing model errors) and its components within a multiple objective framework. MOCOM-UA is a Pareto-based approach that yields an optimal front (or surface) of non-unique solutions in an N-dimensional space, where N is the number of objective-functions. The resulting

Table 2. List of ULM soil parameters from Sac and their plausible ranges.

Parameters	Unit	Description	Plausible Range
UZTWM	mm	Upper zone tension water maximum storage	1.0–300
UZFWM	mm	Upper zone free water maximum storage	1.0–300
UZK	day^{-1}	Upper zone free water lateral depletion rate	0.05–0.75
ZPERC	–	Maximum percolation rate	1.0–350
REXP	–	Exponent of the percolation curve equation	0.0–5.0
LZTWM	mm	Lower zone tension water maximum storage	1.0–500
LZFSM	mm	Lower zone free water supplemental maximum storage	1.0–1000
LZFPM	mm	Lower zone free water primary maximum storage	1.0–1000
LZSK	day^{-1}	Depletion rate of the lower zone supplemental free water storage	0.01–0.8
LZPK	day^{-1}	Depletion rate of the lower zone primary free water storage	0.0001–0.025
PFREE	–	Percolation fraction going directly from upper zone to lower zone free water storages	0.0–0.8
PCTIM	–	Impervious fraction of the ground surface	0.0–0.1
ADIMP	–	Maximum fraction of additional impervious area caused by saturation	0.0–0.45
Physically based parameters (not adjusted)			
Canopy resistance	$\text{s}\,\text{m}^{-1}$		
Maximum snow albedo	%		
Leaf area index	–		
Soil porosity	%		

set of parameters from the Pareto solution defines parameter uncertainty attributable to model structural errors (Vrugt et al., 2003), in which optimizing one objective function trades the performance with other objectives. This is in contrast to methods that either yield a single, unique solution (single metric), or methods that yield a range of results as a function of parameter uncertainty.

In our implementation, the calibrations were first performed on the individual criterion, specifically Q, both ET products, and TWSC to obtain an optimal set of model parameters by minimizing errors in the components of NSE. Next, the same procedure was applied to combinations of these criteria, maximizing their individual NSE, to determine the trade-offs between single- and multi-criteria analyses. Three objective functions were used for all calibrations. These were either the three components of NSE described above (single-criterion calibrations), the individual NSE for two criteria plus the sum of their correlations (two-criteria calibrations), or the NSE of each criterion (three-criteria calibrations). To make the problem more computationally tractable, the initial search of parameter space (i.e. burn-in) was limited to 2000 iterations for all cases. The relative impact of calibrations on model performance with respect to different criteria was further quantified through changes in the relative root-mean-square error (rRMSE) between calibrated and control simulations (described below). This metric provides an additional means for inter-basin comparison, because it is a normalized measure that is (nearly) independent of basin or process scale.

For each basin, the performance of the calibrated model was assessed relative to model performance with default parameters (described in greater detail by Livneh et al., 2011), herein CONTROL. The default parameters are comprised of the Noah LSM land surface characteristics from the National Land Data Assimilation System (NLDAS – Mitchell et al., 2004) and Sac parameters based solely on soil texture (Koren et al., 2003). For the major basins (Sect. 2.1,) we also evaluated the utility of incorporating ET (atmospheric balance and remote sensing) and TWSC as described in Sects. 2.2.1–2.2.3. The tributary catchments are an order of magnitude too small for use of either atmospheric balance ET or GRACE-based TWSC, and hence calibrations for these catchments used Q and ET_{SAT}.

It is also important to note the inherent difference in information content among the variables. Discharge, Q, represents an integrated basin flux measured at a single point. ET represents an areal basin flux, while TWSC represents a change in state over time, measured less frequently than the former quantities. Although each criterion represents a change in volume over time, temporal and spatial averaging were necessary to make them consistent. At large scales, calibration was performed using 0.5° model output averaged to a monthly time step (applicable to Q, TWSC, ET_{SAT}, and ET_{AWB}), whereas small-scale calibration used 1/16° model output averaged to a daily time step (applicable for Q, ET_{SAT}). Parameter values were constrained by the ranges outlined in Table 2. Only the model's soil parameters (from the Sac model), which are generally considered to be conceptual, were modified (Table 2). Physically based quantities, such as vegetation parameters, canopy resistance, greenness, albedo, and rooting depth (all from Noah), were not calibrated.

To further evaluate model performance, an analysis of the variability of hydrologic response in both major basin and tributary streamflows was conducted, followed by an examination of model errors at two selected basins. Three components of model response were examined: the lag-1 autocorrelation (persistence), coefficient of variation (variability), and runoff efficiency (precipitation partitioning). The model's ability to reproduce these observed components quantifies its representation of seasonality and its applicability for hydrologic forecasting under different climate scenarios.

In the final part of the analysis, a subset of the domain was selected to further detail model errors. Examining hydrographs of selected major basins and their tributaries provides an additional means to understand the nature of differences between simulated and observed flows and if it is possible to predict how these errors may propagate within a given region. Lastly, overall uncertainties in the model and observational data are discussed including the manner in which they may affect this study's conclusions.

4 Results and discussion

We present single-criterion calibration results from the major basins first, followed by the tributary single-criterion calibrations. We then present and discuss multi-criteria calibration results for both major and tributary basins. Finally, two regions are selected for a general examination of model errors. In all cases, we considered a single realization of model outputs and observations rather than a range of values that reflect respective uncertainties. This allowed for convenient comparison among simulation/observation pairs. However, it should be emphasized that each time series presented below is one realization from a distribution that would result from combinations of parametric, structural, and observational uncertainties. Similarly, only a single model simulation from each calibration (i.e. Pareto front) was selected with the greatest cumulative skill score across objective functions. This single simulation was used for comparison with other calibrations, rather than comparing the entire sets of model simulations that include the inherent tradeoffs from the Pareto methodology.

4.1 Single-criterion calibrations

We first calibrated ULM using a single-criterion approach based on streamflow simulation errors with the objective functions of the NSE components. Figure 4 shows the results of model calibration to streamflow over major basins. Nearly all basins show calibrated streamflows that follow closely with observations. Notable improvements over a priori values in modeled streamflow were realized over COLO, despite the quantitatively poorer performance compared with other basins (performance statistics presented in the next section). Streamflow simulation errors were noted by other investigators over COLO using the Noah LSM (Xia et al., 2012; Vano et al., 2012) which is relevant to the ULM simulations given its heritage from Noah. Errors were attributed to the significant changes that were made to the Noah canopy parameterizations in its latest official NCEP version (v2.8 – noted by Wei et al., 2012) such as stomatal resistance, seasonal leaf area index (LAI), and root distribution, all of which affect ET and runoff generation. These changes generally improved performance; however, the Colorado basin was an exception that was compensated for here by ULM calibrations that allow for greater soil moisture capacity to store and release large snow melt volumes. For other regions, such as CALI and OHIO, control simulations were fairly skillful at capturing dynamics of seasonal low flows, such that only small improvements were obtained from calibrations. For the remaining regions, runoff ratios were generally too high in the CONTROL simulation, requiring in most cases slight reduction in hydraulic conductivity and increases in moisture holding capacity and permeability parameters.

To quantify the relative uncertainty of the two remote sensing ET products, Fig. 5 compares them with the long-term difference (P-Q) between observed precipitation, P, and streamflow, Q. The underlying assumption in this comparison is that, over a sufficiently long time, the net change in soil moisture storage will become small and the ratio of ET to the difference P-Q will approach unity. It should be noted that the $\mathrm{ET_{AWB}}$ and $\mathrm{ET_{SAT}}$ products are for different periods (1979–2010, and 2001–2010, respectively) and are plotted together to facilitate an initial approximation. In nearly all cases (except RED), $\mathrm{ET_{AWB}}$ is larger than P-Q, corresponding to either a negative change in TWS, or measurement uncertainty. $\mathrm{ET_{SAT}}$ is available for both major and tributary basins, where CALI is the only case with $\mathrm{ET_{SAT}} > P$-Q for a major basin as well as the mean of all of its tributaries. This consistent bias, if not an artifact of estimation error, implies a long-term (2001–2010) loss of terrestrial water storage. Tang et al. (2009) tested this algorithm over northern California and found a slightly high bias in ET compared with ground-based Bowen ratio stations, suggesting that the positive bias seen here could be due in part to the algorithm itself. $\mathrm{ET_{SAT}}$ for all other major basins was slightly less than P-Q, where the means of the respective tributaries were also less than P-Q. The general form of the scatter in Fig. 5 shows increasing $\mathrm{ET_{SAT}}$ negative bias with increasing P-Q, characterized by a pseudo-linear slope of slightly less than one. The mean relative biases on the order of 10–20 % are due either to the $\mathrm{ET_{SAT}}$ algorithm, TWS, observational uncertainty in P and Q, or some combination of these.

The requirement of sub-pixel diversity of VI-Ts in the $\mathrm{ET_{SAT}}$ derivation method is examined in Fig. 6 through a comparison of the long-term residual term, P-Q-$\mathrm{ET_{SAT}}$, and the VI and Ts diversity of each basin. Basin-wide $\mathrm{ET_{SAT}}$ monthly averages (mm month^{-1}) are shown, which were computed from 0.05° pixels (described in Sect. 3.2.1). With the exception of CALI, the large basins have a consistently small residual term and a larger VI-Ts diversity as compared with their tributaries. The bias in Fig. 6 appears to be

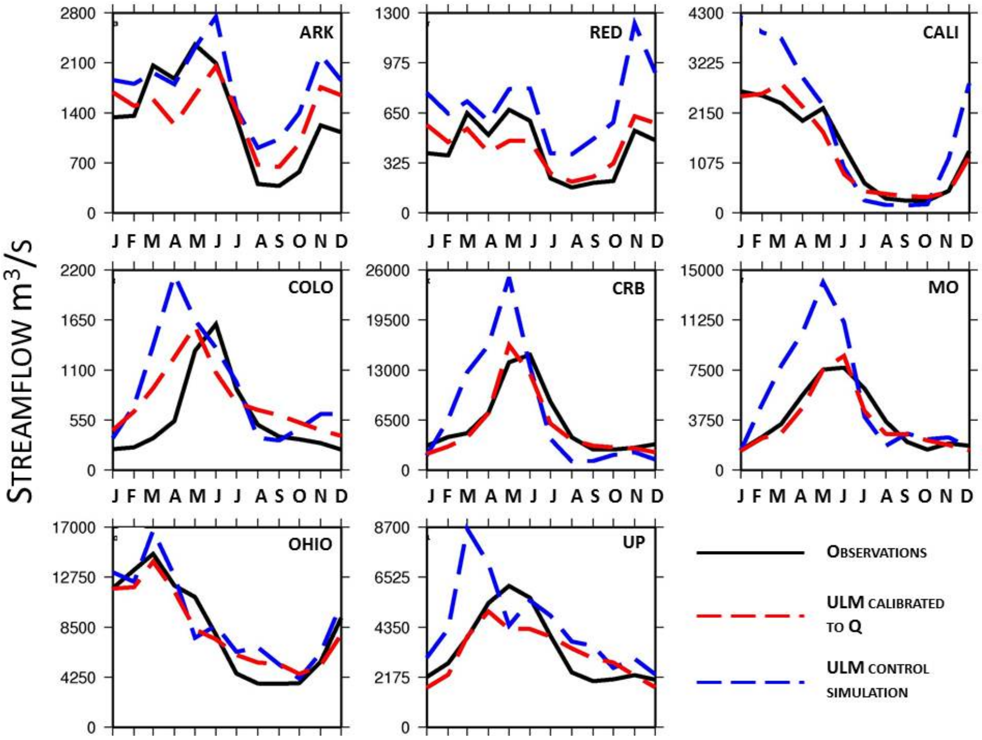

Fig. 4. Mean monthly hydrographs in $m^3\,s^{-1}$ for the major basins for a 20-yr period, the beginning of which varies by basin, depending on data availability (abbreviations defined in Table 1).

irrespective of the VI-Ts diversity, or at minimum does not imply decreasing water balance residual with increasing VI-Ts diversity. Mean NDVI ranges by basin vary from approximately 0.05–0.58, while skin temperature ranges vary from 46–72 K throughout the simulation period. For example, the tributaries of MO possess among the smallest VI-Ts product range, while their water balance residuals are near zero, while basins from CALI have larger VI-Ts diversity products with comparatively larger water balance residual. The implications of Fig. 6 for this analysis are that these basins possess adequate VI-Ts diversity for the ET_{SAT} algorithm. Alternatively stated, the relative VI-Ts diversity alone cannot be used as a means to qualify or disqualify the ET_{SAT} data used here for model calibration.

The two remote sensing ET sources show notable seasonal differences in Fig. 7. For all basins, the ET_{AWB} peaks earlier in the year on average relative to ET_{SAT}, with greater peak magnitude in all cases except CALI. The calibrations were most effective in improving the seasonality and timing of peak ET, whereas calibration improved total ET (monthly) magnitude only for cases where the CONTROL ET was already larger than the respective remote sensing ET product. For cases where either ET_{SAT} or ET_{AWB} were appreciably larger than simulated control ET (most frequently for ET_{AWB}), the calibrated ET remained less than the respective ET product. This difference in ET magnitude was greatest for the westernmost basins, which generally exhibit warm, dry summers with large ET demand. This discrepancy comes about in part because of the constraint imposed by ULM's water balance, something that the remote sensing products do not reflect, and often plays a role when ET demand is high. In these cases, the remote sensing product approaches PET and exceeds the available moisture for actual ET. Over the cold season (DJF), calibrated-ULM frequently matched ET_{SAT}, whereas the larger cold season ET_{AWB} exceeded the calibrated model estimates at all but ARK and LOW, which have comparatively mild cold seasons. Notwithstanding the westernmost basins, the differences between the calibrated model and the respective ET (calibration objective-function) in Fig. 7 are notably less than the difference between the two remote sensing data sets, which can be considered a measure of observational uncertainty.

The seasonal cycle of modeled TWSC has similar amplitude to the GRACE product for most of the basins, as shown in Fig. 8. Whisker plots denote interannual variability of simulations that are not to be confused with model parametric uncertainties. In nearly all cases, calibration brings the mean simulated TWSC within the envelope of observational uncertainty for mean TWSC (denoted by the dark-gray shading). In relative terms, the CALI region has the largest seasonal cycle for both the observed and simulated signals, while regions such as ARK, GBAS, and MO have much smaller amplitudes that are well replicated by ULM. Modest TWSC discrepancies can be expected since we are

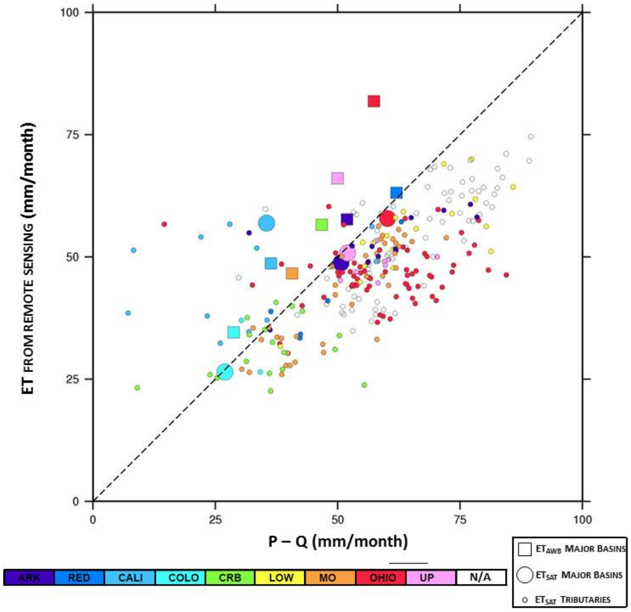

Fig. 5. Estimates of mean monthly evapotranspiration by an atmospheric water balance ($\mathrm{ET_{AWB}}$ – Sect. 2.2.1) in squares, and through satellite data ($\mathrm{ET_{SAT}}$ – Sect. 2.2.2) in circles compared with the residual of precipitation, P, minus streamflow, Q, for the major basins and smaller tributaries (smaller circles). Shaded areas denote the domain within which ET was estimated, such that un-shaded circles represent ET from tributaries outside the major basins.

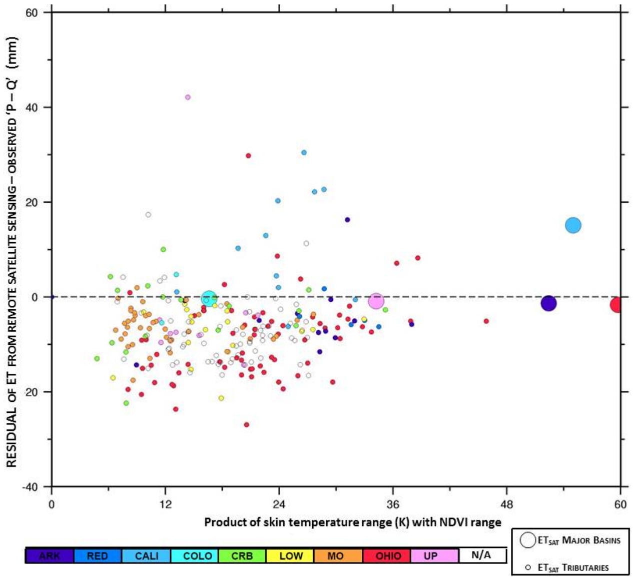

Fig. 6. Comparisons of the residual of evapotranspiration from satellite data ($\mathrm{ET_{SAT}}$ – Sect. 2.2.2) with precipitation, P, minus streamflow, Q, for the major river basins (larger circles) and smaller tributaries (smaller circles) 2001–2010, as a function of VI-Ts diversity, expressed as a product of the ranges of NDVI and skin temperature for each basin. Departures from the dashed line denote either an uncertainty in ET estimates, or significant long-term TWS, or other observational errors.

comparing the model – which is constrained by a relatively shallow ($\sim 2\,\mathrm{m}$) water balance – to the unconstrained estimate of TWSC made by GRACE, which may include contributions from deep groundwater movement and has a coarser native spatial resolution.

The single-criterion calibrations for the tributaries were organized by classifying each catchment by its aridity index, AI, a metric first proposed by Budyko (1974):

$$\mathrm{AI} = R_{\mathrm{net,ann}}/\mathrm{LP}_{\mathrm{ann}}, \tag{4}$$

in which $R_{\text{net-ann}}$ is the annual average net radiation, L is the latent heat of vaporization, and P_{ann} is the mean annual precipitation, such that $\mathrm{LP}_{\mathrm{ann}}$ is the amount of energy needed to evaporate the available precipitation, P_{ann}. AI values exceeding 1 denote increasingly arid (or water limited) conditions, whereas values less than unity denote moist (or radiation limited) conditions. Figure 9 shows the resulting daily calibrated NSE values for the tributaries. Daily NSE values are expected to be smaller than for monthly flows, due to the increased variability in observed flows at the finer temporal scale, which is indeed the case in Fig. 9. A large number of the total tributaries have AI between 0.6 and 1.2. With the exception of two tributaries of RED, the model performance appears to decrease with increasing AI, beginning at AI ≈ 0.6. Figure 10 shows a similar plot but for ET calibrations. Given the seasonality of ET and its strong dependence on atmospheric forcing – i.e. downwelling radiation – many of the tributaries have NSE values above 0.6, with higher NSE values than for the corresponding Q calibrations. However, for a small number of cases (6), ET calibration could not raise model NSE above zero – e.g. less skill than the long-term mean. These disagreements result from cases in the southern part of the domain where $\mathrm{ET_{SAT}}$ values are not constrained by water availability (arid basins) and peak $\mathrm{ET_{SAT}}$ values are in some cases greater than twice the peak modeled values. Notwithstanding specific NSE values for the aforementioned single-criterion calibrations, the degree of improvement resulting from calibration relative to the CONTROL case is presented in greater detail in the following section.

4.2 Multi-criteria calibrations

A central objective of this study was to examine the extent to which calibration towards multiple criteria could improve model simulations relative to each of the criteria. A visual representation of the multi-criteria calibration for the major basins is shown in Fig. 11, while the entire set of results is tabulated in Tables 3 and 4. The three axes in Fig. 11a represent objective functions (NSE) geared towards minimizing modeled errors in, Q, ET, and TWSC, respectively. Within each calibration set, a single optimal solution was selected that represents a tradeoff between optimizing its

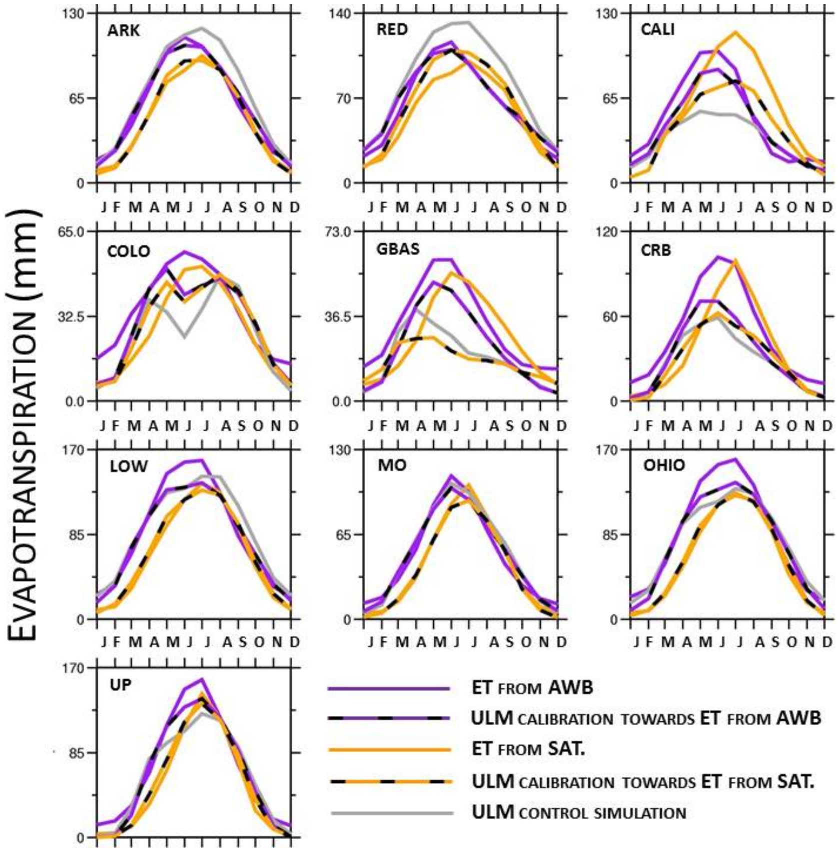

Fig. 7. Mean monthly ET (mm) for the major river basins for the period 2001–2010 that include two sets of calibrations, satellite-based (SAT) or atmospheric water balance-based (AWB) observational products as well as the control simulation.

respective objective functions, giving equal weight to each. Figure 11 is not to be confused with a Pareto front, but rather shows a single optimum simulation selected from individual Pareto fronts for each single- or multi-criteria calibration. As stated in Sect. 3.4, multi-criteria calibrations considered objective functions of the NSE of each criterion, whereas single-criterion calibrations considered objective functions to be the components of the NSE. Consider the example of the calibration labeled Q, $\mathrm{ET_{AWB}}$ that produced a set of simulations that minimized the objective functions for each of these quantities (Q and $\mathrm{ET_{AWB}}$), creating an envelope of similarly scoring simulations (a Pareto front). In order to select the optimal calibration from among these, the simulations that best minimized errors in the auxiliary criterion – in this case TWSC – were chosen. From Fig. 11a it is clear that single-criterion calibrations often lead to poor performance in the other criteria. The exceptions to this pattern are the single-criterion Q-calibrations, which have the largest number of simulations closest to the ideal point (1.0, 1.0, 1.0). Double- and triple-criteria calibrations that include Q were generally the next closest to ideal, with those containing TWSC generally more successful. Conversely, calibrations that did not include Q more frequently performed poorly in one or more criteria, as they lack the implicit overall water balance associated with high fidelity Q simulations – i.e. the timing and partitioning of surface runoff, which encompasses water availabilities for both ET and TWSC. It is assumed here that the observational uncertainties associated with the $\mathrm{ET_{SAT}}$, $\mathrm{ET_{AWB}}$, and TWSC objective functions are larger than for Q observations. Alternatively, poor calibrations may have resulted where those parameters that govern the dominant hydrologic processes were not calibrated. Examples could include regions or seasons where snow melt or vegetation processes were dominant, but parameters controlling those model responses were not optimized (see Table 2).

The extent to which each criterion was improved through calibration is illustrated in Fig. 11b, quantified by the rRMSE difference with each basins CONTROL simulation. Examining this figure along with the accompanying tables (Tables 3, 4), it is clear that calibrations to certain criteria have the potential to either improve or worsen model performance towards other criteria. These tradeoffs should be distinguished from the implied tradeoffs in multi-objective calibrations, since the two are not strictly the same. Figure 11b shows that the results form a central cluster with three branches. The central cluster of simulations is comprised mostly of

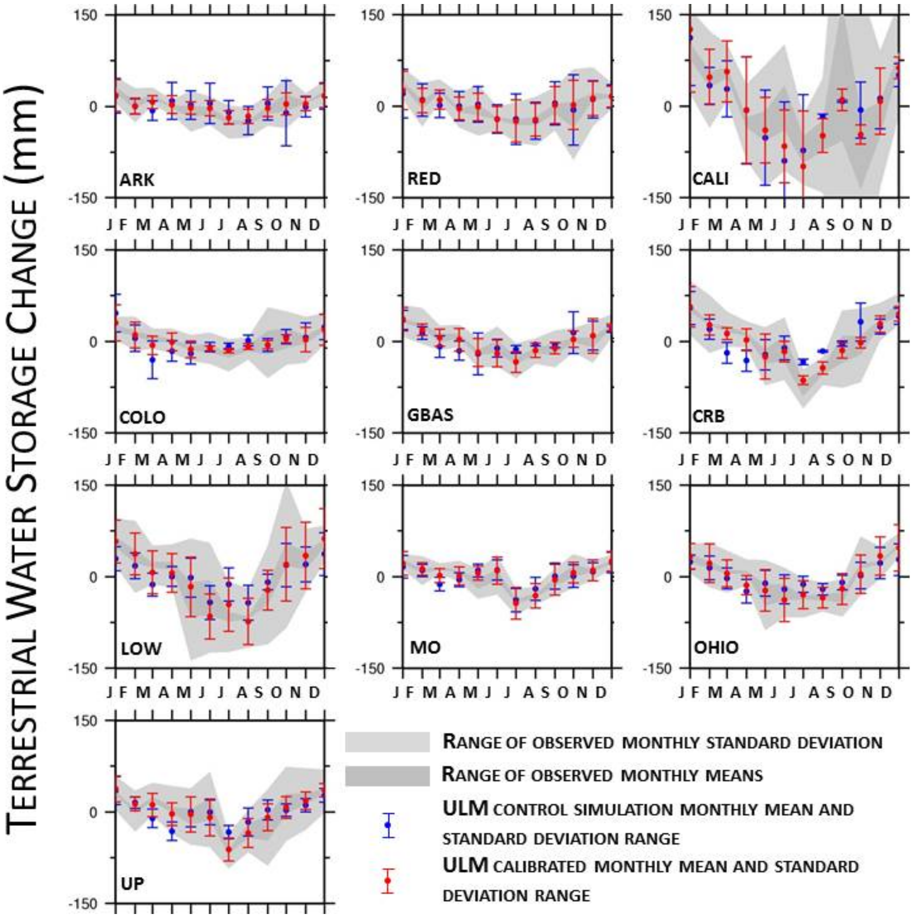

Fig. 8. Mean monthly TWSC (mm) for the major river basins for the period 2002–2010 including the control and calibrated model simulations; the range of variability for each case is shown accordingly.

multi-criteria calibrations that exhibit modest improvements in each criterion. This modest improvement in each criterion is consistent with the degree of improvement noted by Gupta et al. (1999) for their multi-criteria calibrations towards surface heat fluxes. The lower branch is made up mostly of single-criterion ET calibrations ($\mathrm{ET_{SAT}}$, $\mathrm{ET_{AWB}}$) that exclusively improve ET performance, for cases where the objective function conflicts with the other criteria, and hence worsens the performance in other criteria. The upper-left branch is made up of calibrations for which there is good agreement between the ET and TWSC data, and hence large improvements in these objective functions through calibration. The upper-right branch follows similarly except with agreements between TWSC and Q data.

This analysis suggests that, at the regional scale (larger than $\approx 10^5\ \mathrm{km}^2$), calibrations towards Q are generally more robust than those towards TWSC and ET in a multi-criteria context. Overall, the remote-sensing auxiliary criteria ($\mathrm{ET_{AWB}}$, $\mathrm{ET_{SAT}}$, TWSC) generally provide useful information regarding the seasonality of the terrestrial water balance. However, these criteria alone or in combination do not appear sufficient to appreciably improve model simulations of Q, as may be the desire in an ungauged basin.

Figure 12 shows multi-criteria results for the tributaries and follows the same format as Fig. 11 with considerably more data points for the Q and $\mathrm{ET_{SAT}}$ criteria. In contrast to the major basin results in Fig. 11a, Fig. 12a shows that the multi-criteria calibration (Q, $\mathrm{ET_{SAT}}$) for the tributaries performs competitively with both single-criterion calibrations in terms of NSE for a large number of tributaries. For all calibration criteria, there are basins that perform poorer than climatology – i.e. NSE < 0. However, these are mostly for single-criterion calibrations relative to the other criterion. For example, it follows intuitively that the $\mathrm{ET_{SAT}}$ calibration has instances of poorer NSE with respect to Q than does the Q, $\mathrm{ET_{SAT}}$ calibration. Figure 12b shows quantitatively greater improvements in Q performance than ET (note that the horizontal axes are not the same in this plot). This reflects the greater flexibility in model structure and (soil) parameter combinations considered here to influence Q outputs versus ET with relation to a given set of atmospheric forcings. For both Q and $\mathrm{ET_{SAT}}$, rRMSE improvements in

Table 3. Summary of skill scores and improvements with respect to model a priori performance, from the single-criterion calibrations; numeric values show improvement, while dash cells indicate no improvement in model skill for the respective variable. Underlined values denote the specific ET observation to which calibration was performed.

Calibration quantity		NSE skill				rRMSE improvement			
	Basin	Q	ET_{AWB}	ET_{SAT}	TWSC	Q	ET_{AWB}	ET_{SAT}	TWSC
Q	ARK	0.85	0.88	0.89	0.17	1.68	0.06	0.18	2.88
	RED	0.78	0.81	0.70	0.37	2.65	0.15	0.19	2.03
	CALI	0.94	0.75	0.10	0.53	0.48	0.20	0.02	4.19
	COLO	0.46	0.57	0.50	0.32	5.78	0.14	0.11	6.01
	CRB	0.78	0.50	0.63	0.69	0.77	0.06	0.14	7.30
	MO	0.87	0.77	0.74	0.61	1.76	–	–	11.49
	OHIO	0.86	0.68	0.76	0.43	0.21	–	0.02	0.80
	UP	0.72	0.54	0.56	0.20	0.46	–	–	2.60
ET_{AWB}	ARK	–	0.93	0.76	0.03	–	0.11	0.08	–
	RED	–	0.89	0.62	0.39	–	0.21	0.15	2.66
	CALI	0.22	0.84	0.36	0.41	–	0.27	0.11	2.17
	COLO	–	0.61	0.56	0.05	3.65	0.15	0.13	2.63
	GBAS	–	0.62	0.43	0.57	–	0.21	0.23	3.23
	CRB	–	0.63	0.54	0.67	–	0.13	0.08	6.96
	LOW	–	0.92	0.76	0.42	0.00	0.05	0.04	6.17
	MO	–	0.93	0.78	0.47	0.36	0.02	–	6.92
	OHIO	0.15	0.92	0.74	0.37	–	0.05	0.00	–
	UP	–	0.93	0.82	0.04	–	0.10	–	–
ET_{SAT}	ARK	–	0.81	0.96	–	–	–	0.27	–
	RED	–	0.77	0.90	0.32	–	0.12	0.32	0.62
	CALI	0.45	0.45	0.65	0.31	–	0.03	0.24	0.75
	COLO	–	0.53	0.69	0.02	3.16	0.12	0.20	2.26
	GBAS	–	0.59	0.53	0.60	–	0.20	0.27	3.57
	CRB	–	0.45	0.65	0.27	0.05	0.04	0.15	1.43
	LOW	–	0.73	0.97	0.20	0.15	–	0.26	1.62
	MO	–	0.78	0.96	0.20	–	–	0.22	–
	OHIO	–	0.71	0.96	0.32	–	–	0.25	–
	UP	–	0.84	0.96	0.06	–	–	0.21	–
TWSC	ARK	–	0.80	0.87	0.37	–	–	0.16	11.49
	RED	–	0.55	0.24	0.57	–	0.02	0.00	8.38
	CALI	0.15	0.71	0.07	0.47	–	0.17	0.00	3.22
	COLO	–	0.11	0.26	0.19	2.53	0.00	0.01	4.25
	GBAS	–	0.59	0.48	0.59	–	0.20	0.25	3.52
	CRB	–	0.53	0.54	0.68	–	0.07	0.07	7.17
	LOW	–	0.77	0.63	0.58	–	–	–	9.97
	MO	–	0.91	0.82	0.52	0.52	–	0.02	8.49
	OHIO	0.39	0.69	0.54	0.55	–	–	–	2.70
	UP	–	0.75	0.76	0.42	0.13	–	–	7.75

single-criterion calibrations were frequently made at the expense of rRMSE of the other criterion. An interesting finding is that the top Q simulations from approximately one-third of all tributaries (81) resulted from multi-criteria Q, ET_{SAT} calibrations. This is in direct contrast to the major basin calibrations, in which the top performing Q simulations resulted exclusively from single-criterion Q calibrations. For only six tributaries ($\sim 2\,\%$ of all tributaries), ET_{SAT} calibrations improved Q to a comparable degree to Q calibrations. Therefore, consistent with the major basin analysis, the use of only auxiliary remote-sensing criteria (in this case, only ET_{SAT}) was not sufficient to appreciably and reliably improve Q performance. The unique conclusion here is that the inclusion of an auxiliary remote sensing criterion (Q, ET_{SAT}) for the tributary basins ($< 10^4\,\mathrm{km}^2$) leads to improved calibration results *beyond* that of the single-criterion calibration. This suggests that the initial single-criterion calibration did not find the global optimum. Given the constrained burn-in

Table 4. Same as Table 3, except for multi-criteria calibrations.

Calibration quantity		NSE skill				rRMSE improvement			
	Basin	Q	ET_{AWB}	ET_{SAT}	TWSC	Q	ET_{AWB}	ET_{SAT}	TWSC
QET_{AWB}	ARK	0.59	0.89	0.74	–	1.43	0.07	0.06	–
	RED	0.09	0.81	0.57	–	2.14	0.15	0.13	–
	CALI	0.68	0.64	0.27	0.40	0.13	0.13	0.08	2.04
	COLO	–	0.55	0.55	0.09	4.11	0.13	0.13	3.06
	CRB	–	0.16	0.26	0.41	0.21	–	–	3.10
	MO	–	0.87	0.77	0.29	1.24	–	–	2.06
	OHIO	0.74	0.89	0.76	0.47	0.12	0.01	0.02	1.40
	UP	0.71	0.86	0.81	0.31	0.45	0.02	–	5.18
QET_{SAT}	ARK	0.50	0.87	0.75	–	1.36	0.04	0.07	–
	RED	–	0.74	0.54	–	2.02	0.10	0.11	–
	CALI	0.71	0.45	0.62	0.35	0.15	0.03	0.22	1.29
	COLO	–	0.55	0.69	0.02	3.95	0.13	0.20	2.26
	CRB	–	0.19	0.58	0.42	0.35	–	0.10	3.31
	MO	–	0.86	0.72	0.22	1.25	–	–	0.25
	OHIO	0.69	0.78	0.77	0.38	0.09	–	0.03	0.04
	UP	0.60	0.88	0.82	0.23	0.39	0.04	–	3.35
Qtwsc	ARK	0.46	0.87	0.68	0.18	1.35	0.04	0.03	0.68
	RED	–	0.82	0.66	0.45	–	0.15	0.17	4.34
	CALI	0.71	–	–	0.38	0.17	–	–	1.81
	COLO	–	0.53	0.60	0.05	4.21	0.13	0.15	2.65
	CRB	0.09	–	0.01	0.40	0.39	–	–	3.04
	MO	0.06	0.87	0.75	0.27	1.30	–	–	1.67
	OHIO	0.73	0.85	0.71	0.50	0.11	–	–	1.82
	UP	0.55	0.63	0.65	0.32	0.37	–	–	5.45
ET_{AWB}TWSC	ARK	–	0.88	0.86	0.35	–	0.06	0.15	10.13
	RED	–	0.59	0.32	0.49	–	0.03	0.03	5.46
	CALI	0.15	0.76	0.11	0.43	–	0.21	0.02	2.47
	COLO	–	0.64	0.63	0.14	3.94	0.17	0.17	3.68
	GBAS	–	0.59	0.47	0.59	–	0.20	0.25	3.50
	CRB	–	0.56	0.56	0.69	–	0.09	0.09	7.30
	LOW	–	0.85	0.71	0.58	–	–	0.00	9.92
	MO	–	0.93	0.79	0.50	0.66	0.02	–	7.95
	OHIO	0.42	0.90	0.71	0.54	–	0.02	–	2.46
	UP	–	0.88	0.80	0.37	0.13	0.04	–	6.66

of 2000 iterations, the single-criterion optimization of Q was likely clustered about a local optima, which has been noted by others using the MOCOM-UA procedure (Vrugt et al., 2003). Hence, we conclude that it was the additional signal (in this case, ET_{SAT}) that forced the parameter search towards more optimal areas within the parameter space.

4.3 Hydrologic response and model error analysis

Calibrated model parameters for this extended streamflow analysis were selected from Sect. 4.2 based on the best performing Q calibrations. In the case where several of the best calibrations have similar skill in simulating Q (arbitrarily NSE values within 5 % of one another), the parameters associated with the simulation with higher performance in the auxiliary criteria were selected – i.e. ET_{SAT}, ET_{AWB}, and TWSC for major basins, ET_{SAT} only for tributaries. As part of this validation, basins were screened for a period of record that was considerably longer than the calibration window (18 yr), chosen here to be ~ 70 yr, to provide a robust characterization of their hydrologic response.

Table 5 shows the simulated and observed runoff efficiencies, lag-1 autocorrelations, and coefficients of variation for both major basins and tributaries. These variability components were computed using flows at a monthly time scale to facilitate direct comparison between major and tributary flow responses, since most major basin streamflows were only available monthly for the cases of naturalized flows. Runoff efficiencies were fairly well matched by ULM across basins and scales, with a few exceptions, most notably COLO. For

Table 4. Continued.

Calibration quantity		NSE skill				rRMSE improvement			
	Basin	Q	ET_{AWB}	ET_{SAT}	TWSC	Q	ET_{AWB}	ET_{SAT}	TWSC
ET_{SAT}TWSC	ARK	–	0.89	0.89	0.32	–	0.05	0.18	8.08
	RED	–	0.79	0.79	0.43	–	0.13	0.24	3.62
	CALI	0.52	0.70	0.70	0.38	0.00	0.20	0.27	1.81
	COLO	–	0.71	0.71	0.10	3.31	0.12	0.21	3.23
	GBAS	–	0.50	0.50	0.59	–	0.12	0.26	3.53
	CRB	–	0.40	0.50	0.30	0.94	0.16	0.18	2.54
	LOW	–	0.77	0.77	0.52	–	–	0.05	8.40
	MO	–	0.82	0.82	0.49	0.55	–	0.02	7.65
	OHIO	0.39	0.79	0.79	0.48	–	–	0.04	1.46
	UP	0.42	0.81	0.81	0.38	0.32	–	–	6.88
QET_{AWB}TWSC	ARK	0.43	0.89	0.71	–	1.33	0.06	0.04	–
	RED	–	0.80	0.47	–	2.04	0.14	0.09	–
	CALI	0.63	0.64	0.41	0.37	0.09	0.13	0.13	1.59
	COLO	–	0.53	0.60	0.05	4.21	0.13	0.15	2.65
	CRB	–	0.11	0.39	0.63	0.31	–	–	6.36
	MO	–	0.92	0.80	0.40	0.54	0.11	0.14	1.83
	OHIO	0.73	0.88	0.72	0.51	0.11	0.01	–	1.95
	UP	0.53	0.77	0.77	0.27	0.36	–	–	4.13
QET_{SAT}TWSC	ARK	0.49	0.73	0.73	–	1.37	–	0.06	–
	-RED	-	0.46	0.46	–	1.97	–	0.08	–
	CALI	0.63	0.41	0.41	0.37	0.09	0.06	0.13	1.59
	COLO	–	0.60	0.60	0.05	4.21	0.06	0.15	2.65
	CRB	0.09	0.01	0.01	0.40	0.39	–	–	3.04
	MO	–	0.80	0.80	0.30	1.04	0.11	0.09	2.83
	OHIO	0.74	0.74	0.74	0.45	0.12	–	0.01	1.12
	UP	0.63	0.79	0.79	0.34	0.41	–	–	5.76

Table 5. Variability analysis for observed followed simulated trends by major basin and tributary averages over a 70-yr period, including runoff efficiency, R_e, lag-1 autocorrelation, r_1, and the coefficient of variation, CV.

	Major Basin						Total sub-basins*	Sub-basin tributary averages					
	R_e		r_1		CV			R_e		r_1		CV	
	obs.	sim.	obs.	sim.	obs.	sim.		obs.	sim.	obs.	sim.	obs.	sim.
ARK	0.14	0.15	0.49	0.49	1.03	1.01	9/12	0.21	0.24	0.35	0.52	1.97	2.12
RED	0.12	0.14	0.46	0.48	1.14	1.02	4/5	0.16	0.15	0.52	0.77	1.38	1.61
CALI	0.46	0.45	0.65	0.65	1.02	1.05	7/11	0.41	0.49	0.60	0.69	1.77	1.39
COLO	0.10	0.14	0.68	0.72	0.99	0.92	2/2	0.41	0.67	0.55	0.68	1.39	1.14
CRB	0.45	0.45	0.72	0.66	0.80	0.85	9/18	0.36	0.41	0.63	0.64	1.11	1.09
MO	0.12	0.14	0.67	0.79	0.72	0.78	13/41	0.20	0.25	0.46	0.57	1.56	1.35
OHIO	0.41	0.40	0.63	0.71	0.74	0.60	46/66	0.41	0.45	0.48	0.60	0.95	0.75
UP	0.27	0.28	0.74	0.67	0.67	0.60	8/9	0.28	0.27	0.49	0.60	1.13	0.89
LOW	NA	NA	NA	NA	NA	NA	11/18	0.27	0.25	0.45	0.64	1.32	1.33
Other	NA	NA	NA	NA	NA	NA	54/68	0.34	0.39	0.48	0.58	1.05	1.05

*The number of tributaries was reduced to the first number from the second number with the requirement for ∼ 70 yr flow record.

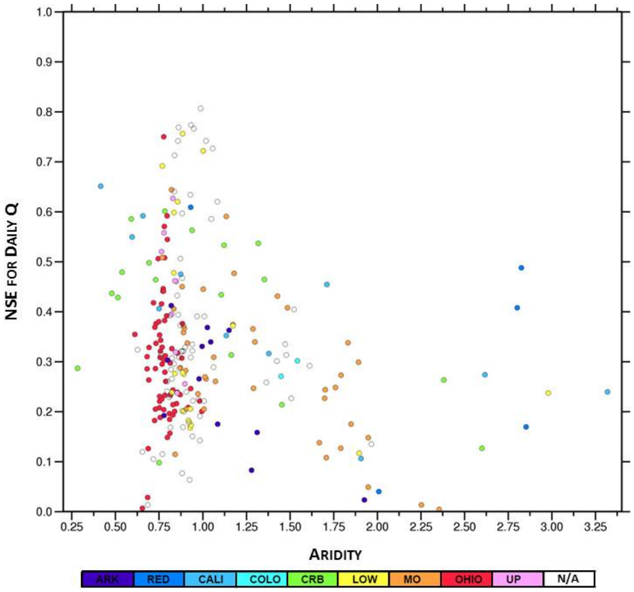

Fig. 9. NSE values for ULM calibrations to streamflow at a daily time step as a function of AI for the period 1991–2010. Shading of individual points denotes the major region for each tributary.

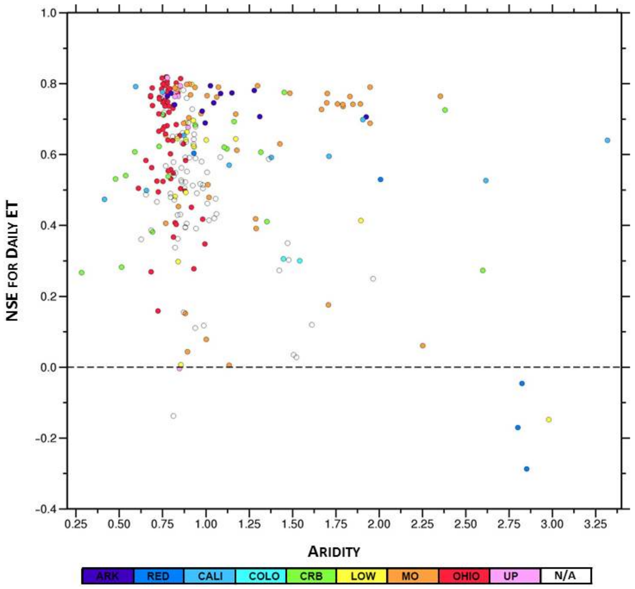

Fig. 10. NSE values for ULM calibrations towards ET_{SAT} at a daily time step as a function of AI for the period 2001–2010. Shading of individual points denotes the major region for each tributary.

cases of large runoff efficiency discrepancy – i.e. larger than 10 % – simulated Nash-Sutcliffe efficiencies were consistently higher than observed. This could result from model errors such as negative biases in ET estimates (noted for several basins in Sect. 4.1), inadequate soil moisture storage capacity, or negative biases in the precipitation forcing, all of which could produce higher runoff efficiency than observed. Model persistence (i.e. lag-1 autocorrelation) follows observations reasonably well. For cases of notable disagreement, simulated persistence was most frequently higher than observed, which may be due in part to a lack of information of extreme/localized meteorological events in the forcing data. The major basins UP and CRB are unique in this regard, where the model is less persistent than observations. Persistence errors do not appear to be related to coefficient of variation errors, as modeled CV was both higher and lower than observations for basins where modeled flows were more persistent than observations. Modeled CV values were the most varied and did not show a systematic bias across basins or across scale.

Two major basins were selected to examine streamflow errors more closely: the CRB and OHIO. To enable a visual comparison among basins with different flow magnitudes, the streamflows in Fig. 13 were converted to z-scores, via subtraction of the long-term observed mean flow and division by the standard deviation. CRB has a variety of interesting hydroclimatic features such as alpine, maritime and arid regions, and its tributaries possess the widest range of AI values of any region. Model errors for the major basin are concentrated near the time of peak flow, relating to snowmelt dynamics in this heavily snowmelt-influenced region. The major basin model flows were less persistent with higher CV than observations, which is consistent with the sharper peak in the hydrograph. The time of peak flow comes, on average, one month earlier in the tributaries, reflecting their rapid response and shorter times of concentration. Over the tributaries, the model tends to under (over) predict high (low) flows, such that beginning at the time of peak flow tributary errors tend to precede major basin errors by approximately one month. The ranges of AI values and snow versus rain-dominated conditions between the major basin and its tributaries are depicted in the multiple hydrographs of the bottom panel, revealing that the large snowmelt-dominated component of the major hydrograph was only scarcely sampled by the tributaries in the study domain.

The Ohio River Basin is situated within a more uniform continental hydroclimatic regime than the CRB. Its tributaries are more numerous (46, versus 9 in CRB) and hence represent an even more comprehensive range of conditions specific to the region. The timing of maximum and minimum flows is remarkably similar between the mean of the tributaries and the major flows, consistent with the relative hydroclimatic homogeneity of the region. Similar to CRB, the model under (over) predicted high (low) flows, while high correlation among tributary streamflow is evident from their respective hydrographs. The smoother simulated hydrographs at both scales is consistent with the overestimated persistence and underestimated variance noted in Table 5.

Given the multiple data sets used in this study, it is essential to temper the findings with the impact of overall data uncertainty. To train the model, several independent data sets were used that could lead to offsetting errors across these

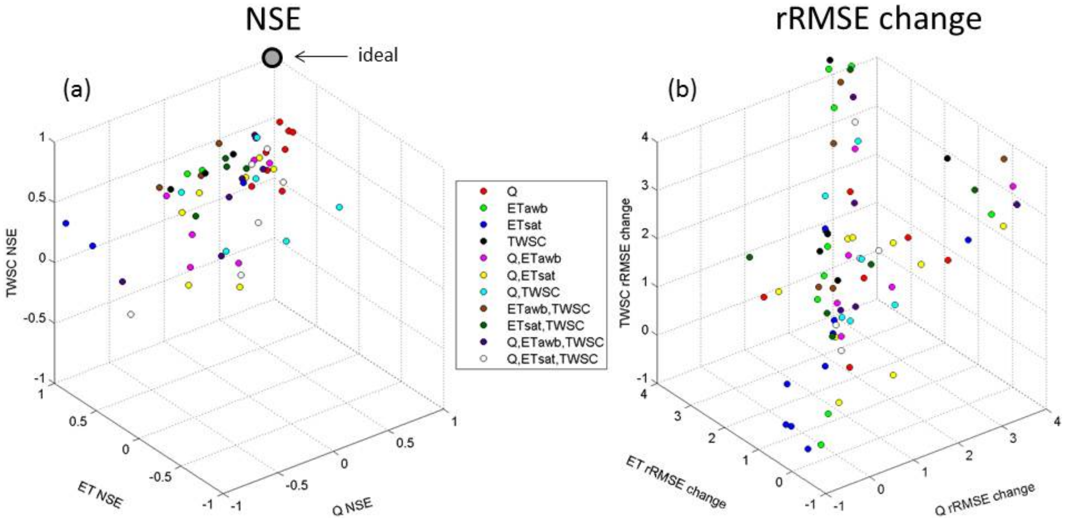

Fig. 11. ULM calibrations over major basins towards combinations of Q, $\mathrm{ET}_{\mathrm{SAT}}$, ETAWB, and TWSC at a monthly time step for the period 1991–2010, including **(a)** NSE values for each criterion (cutoff at -1 for clarity), and **(b)** differences in rRMSE for each criterion resulting from the respective calibrations. The entire set of results for these plots is included in Table 3.

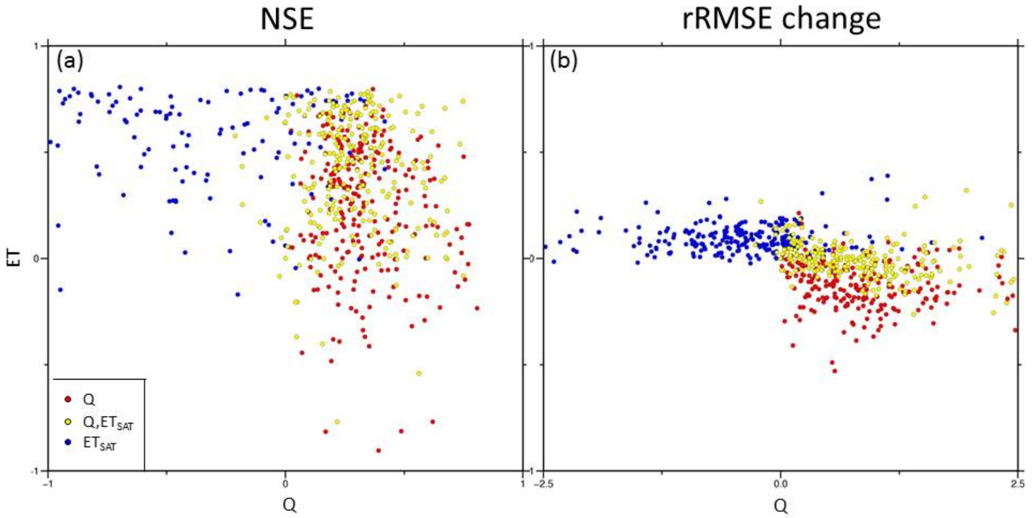

Fig. 12. ULM calibrations over tributary basins towards combinations of Q, and $\mathrm{ET}_{\mathrm{SAT}}$ at a daily time step for the period 1991–2010, including **(a)** NSE values for each criteria and **(b)** differences in rRMSE for each criteria resulting from the respective calibrations.

data sets – i.e. ET versus TWSC – highlighting potential water budget inconsistencies and data uncertainties. This is an inherent potential pitfall in using independent data sets; however, it may aid in ultimately bracketing true conditions. One technique to reconcile such inconsistencies is through redistributing the total water balance error from multiple sensors back to each of the individual components using a Kalman error approach (Pan et al., 2012). This approach is beyond the scope of this work; however, it may offer a framework to further improve the consistency of the remote sensing water budget analyses in the future.

Overall sources of error and uncertainty are as follows. TWSC uncertainties were perhaps the largest within the study, evident in the disparity between mean monthly values from individual processing streams in Fig. 8. A further uncertainty arose in comparing these data with ULM, given the different reference depths considered by each. $\mathrm{ET}_{\mathrm{AWB}}$ errors were most likely to arise from the atmospheric components (left side of Eq. 1) that were contingent upon the NARR analysis increment, particularly problematic over coastal regions (Ruane, 2010), wherein adjustments to latent heating of the atmospheric column are made to overcome moisture excesses in the underlying Eta model. $\mathrm{ET}_{\mathrm{SAT}}$ estimates were subject to uncertainty from the input MODIS skin temperatures, which Ferguson et al. (2010) described as being the largest source of error for a similar satellite-based ET

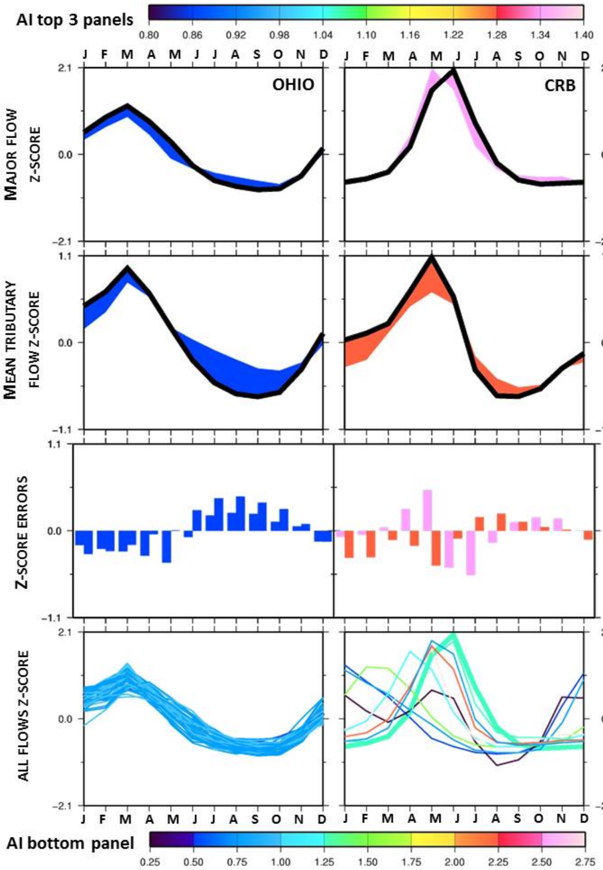

Fig. 13. Comparison between major basin and mean tributary flows and errors. Flow data was converted to z-scores to allow for comparison among basins. The differences in aridity index (AI) are shaded according to the upper-scale for the top 3 panels and the lower scale for the bottom panel.

product, in lieu of errors associated with emissivity and land-surface characteristics. Furthermore, the $\mathrm{ET_{SAT}}$ estimates are not strictly constrained by moisture availability that could lead to further data uncertainty. Q uncertainties due to random errors in the USGS current meters or (for major basins) the different naturalization algorithms could skew the interpretation of the model performance and trend analysis. Albeit, it is not immediately apparent in which direction the skew would occur and it is expected – at least in the case of the in-situ USGS data – that these errors would be small relative to the greater complexities in the remote sensing data. Meteorological forcing errors may exist, particularly in regions of topographical complexity. Most notably, precipitation errors would prevent the model from matching streamflow timing, magnitude, and variability, while surface temperature and wind errors could translate into erroneous estimates of surface water and energy fluxes. Lastly, errors in model structure or conceptualization errors may exist that ultimately prevent the model from correctly simulating certain processes, or achieving the correct results for the incorrect reason via calibration. Investigating these types of errors would require a more directed and rigorous error analysis including plot-scale evaluation of soil and vegetation parameterizations (i.e. alternative model structures), evaluation of land-cover inputs, in addition to comparisons with detailed measurements of surface fluxes of moisture and energy and their respective uncertainties. Clearly, COLO was sensitive to these types of errors, and it is expected that other arid regions, as well as regions subject to climatic extremes, would exhibit large sensitivities.

5 Conclusions

We exploited several observational data sets together with an LSM to estimate various components of the terrestrial water budget. The analysis focused on ways to train ULM to observational data sets to improve estimates of water budget components. The results were presented to provide insight into tradeoffs in the performance with respect to each criterion. The single best performing streamflow parameters for each basin were utilized to assess streamflow variability and hydrologic response. Finally, an examination into potential error sources was made to illustrate specific causes behind discrepancies in simulated streamflows and their relationship across scale. The most important conclusions of this analysis are the following:

1. Model calibrations towards a single criterion had varied results. Over major basins ($\geq 10^5\,\mathrm{km}^2$), the model was able to replicate Q, ET, and TWSC individually with reasonable skill, despite uncertainties in the data themselves and discrepancies between modeled and native retrieval resolutions. Over 250 tributary-scale basins ($< 10^4\,\mathrm{km}^2$) over daily time steps, ET calibrations generally scored higher than Q calibrations. However, for a small number of these (arid basins), strong disagreements between the model and remote-sensing product lead to ET simulations that were poorer predictors than climatology, while Q calibrations always provided additional skill.

2. Over major basins, calibrations towards multiple-criteria had the best overall performance when Q was included, followed by $\mathrm{ET_{SAT}}$, $\mathrm{ET_{AWB}}$, and TWSC. Altogether, calibrations towards Q alone had the best all-around performance in terms of the other criteria, while neither the other criteria (ET, TWSC) alone nor in combination were able to add appreciable skill to Q prediction, since this would be desirable for training a model in ungauged basins.

3. Multi-criteria performance over tributary-scale basins followed similarly to the large-scale analysis with the notable exception that the multi-criteria calibration (Q and $\mathrm{ET_{SAT}}$ together) out-performed the single-criterion

Q-calibration in terms of Q performance at roughly one-third of the basins. This suggests that traditional streamflow calibration stands to benefit from the inclusion of remote-sensing data.

4. The lack of a systematic bias in the satellite-ET product over a number of basins of varying VI and Ts diversity indicates that, above a certain threshold, VI-Ts diversity alone may not be an adequate predictor of quality of the satellite-based ET product. Rather, the issue of unbounded ET estimates during summer was most detrimental to the quality of ET estimates.

5. The use of multiple criteria in the calibration procedure at minimum serves to reduce the equifinality problem when choosing the "best" instance of the model parameters.

6. Investigating model error sources revealed that simulations generally underpredicted high flows and overpredicted low flows.

Acknowledgements. This work was facilitated through the use of advanced computational, storage, and networking infrastructure provided by the Hyak supercomputer system, supported in part by the University of Washington eScience Institute. This work was supported in part by NOAA Cooperative Agreements No NA08OAR4320899 and NA10OAR4310245 with the University of Washington.

Edited by: B. Schaefli

References

Beven, K. J. and Binley, A.: The future of distributed models – model calibration and uncertainty prediction, Hydrol. Process., 6, 279–298, 1992.

Beven, K. J. and Freer J.: Equifinality, data assimilation, and uncertainty estimation in mechanistic modeling of complex environmental systems using the GLUE methodology, J. Hydrol., 249, 11–29, 2001.

Birkel, C., Tetzlaff, D., Dunn, S. M., and Soulsby, C.: Towards simple dynamic process conceptualization in rainfall-runoff models using multi-criteria calibration and tracers in temperate, upland catchments, Hydrol. Process., 24, 260–275, 2010.

Brocca, L., Melone, F., Moramarco, T., Wagner, W., Naeimi, V., Bartalis, Z., and Hasenauer, S.: Improving runoff prediction through the assimilation of the ASCAT soil moisture product, Hydrol. Earth Syst. Sci., 14, 1881–1893, doi:10.5194/hess-14-1881-2010, 2010.

Budyko, M. I: Climate and Life, Academic Press, New York, 508, 1974.

Burnash, R. J. C., Ferral, R. L., and McGuire, R. A.: A generalized streamflow simulation system - Conceptual modeling for digital computers, Technical Report, Joint Federal and State River Forecast Center, US National Weather Service and California Department of Water Resources, Sacramento, 204 pp., 1973.

Crow, W. T., Wood, E. F., and Pan, M.: Multiobjective calibration of land surface model evapotranspiration predictions using streamflow observations and spaceborne surface radiometric temperature retrievals, J. Geophys Res., 108, 4725, doi:10.1029/2002JD003292, 2003.

Ek, M. B., Mitchell, K. E., Lin, Y., Rogers, E., Grunmann, P., Koren, V., Gayno, G., and Tarpley, J. D.: Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model, J. Geophys. Res., D22, 8851, doi:10.1029/2002JD003296, 2003.

Falcone J. A., Carlisle, D. M., Wolock, D. M., and Meador, M. R.: GAGES: A stream gage database for evaluating natural and altered flow conditions in the conterminous United States, available at: http://www.esapubs.org/archive/ecol/E091/045/, Ecology, 91, 621, 2010.

Ferguson, C. R., Sheffield, J., Wood, E. F., and Gao, H.: Quantifying uncertainty in a remote sensing based estimate of evapotranspiration over the continental United States, Int. J. Remote Sens., 31, 3821–3865, 2010.

Gupta, H. V., Bastidas, L. A., Sorooshian, S., Shuttleworth, W. J., and Yang, Z. L.: Parameter estimation of a land surface scheme using multicriteria methods, J. Geophys. Res., 104, 491–503, 1999.

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

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

João, E.: How scale affects environmental impact assessment, Environ. Impact Asses., 22, 289–310, 2002.

Kalma, J. D., McVicar, T. R., and McCabe, M. F.: Estimating land surface evaporation: a review of methods using remotely sensed surface temperature data, Surv. Geophys., 29, 421–469, 2008.

Kalnay, E., Kanamitsu, M., Kistler, R., Collings, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K. C., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, R., and Josep, D.: The NCEP/NCAR 40-Year Reanalysis Project, B. Am. Meteorol. Soc., 77, 437–471, 1996.

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

Kingston, D. G., Hannah, D. M., Lawler, D. M., and McGregor, G. R.: Regional classification, variability, and trends of northern North Atlantic river flow, Hydrol. Process., 25, 1021–1033, 2011.

Klees, R., Liu, X., Wittwer, T., Gunter, B. C., Revtova, E. A., Tenzer, R., Ditmar, P., Winsemius, H. C., and Savenije, H. H. G.: A Comparison of Global and Regional GRACE Models for Land Hydrology, Surv. Geophys., 29, 335–359, doi:10.1007/s10712-008-9049-8, 2008.

Koren, V. I., Smith, M., and Duan, Q.: Use of *a priori* parameter estimates in the derivation of spatially consistent parameter sets of rainfall-runoff models, in: Calibration of Watershed Models, Water Science and Applications, Vol. 6, edited by: Duan, Q., Sorooshian, S., Gupta, H., Rosseau, H., and Turcotte, H., AGU, 239–254, 2003.

Lischeid, G.: Combining hydrometric and hydrochemical data sets for investigating runoff generation processes: tautologies, inconsistencies and possible explanations, Geography Compass, 2 255–280, doi:10.1111/j.1749-8198.2007.00082.x, 2008.

Livneh, B., Xia, Y., Mitchell, K. E., Ek, M. B., and Lettenmaier, D. P.: Noah LSM Snow Model Diagnostics and Enhancements, J. Hydrometeorol., 11, 721–738, 2010.

Livneh, B., Restrepo, P. J., and Lettenmaier, D. P.: Development of a Unified Land Model for prediction of surface hydrology and land-atmosphere interactions, Journal of Hydrometeorology, 12, 1299–1320, doi:10.1175/2011JHM1361.1, 2011.

Livneh, B., Rosenberg, E. A., Lin, C., Mishra, V., Andreadis, K. M., and Lettenmaier, D. P.: Extension and spatial refinement of a long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States, J. Climate, submitted, 2012.

Lo, M., Famiglietti, J. S., Yeh, P.J.-F., and Syed, T. H.: Improving parameter estimation and water table depth simulation in a land surface model using GRACE water storage and estimated base flow data, Water Resour. Res., 46, W05517, doi:10.1029/2009WR007855, 2010.

MacLean, A. J., Tolson, B. A., Seglenieks, F. R., and Soulis, E.: Multiobjective calibration of the MESH hydrological model on the Reynolds Creek Experimental Watershed, Hydrol. Earth Syst. Sci. Discuss., 7, 2121–2155, doi:10.5194/hessd-7-2121-2010, 2010.

McCabe, M. F., Franks, S. W., and Kalma, J. D.: Calibration of a land surface model using multiple datasets, J. Hydrolog., 302, 209–222, 2005.

Mesinger, F., DiMego, G., Kalnay, E., Mitchell, K., Shafran, P. C., Ebisuzaki, W., Jovic, D., Woollen, J., Rogers, E., Berbery, E. H., Ek, M.B ., Fan, Y., Grumbine, R., Higgins, W., Li, H., Lin, Y., Manikin, G., Parrish, D., and Shi, W.: North American Regional Reanalysis, B. Am. Meteorol. Soc., 87, 343–360, 2006.

Milzow, C., Krogh, P. E., and Bauer-Gottwein, P.: Combining satellite radar altimetry, SAR surface soil moisture and GRACE total storage changes for hydrological model calibration in a large poorly gauged catchment, Hydrol. Earth Syst. Sci., 15, 1729–1743, doi:10.5194/hess-15-1729-2011, 2011.

Mitchell, K. E., Lohmann, D., Houser, P. R., Wood, E. F., Schaake, J. C., Robock, A., Cosgrove, B. A., Sheffield, J., Duan, Q., Luo, L., Higgins, R. W., Pinker, R. T., Tarpley, J. D., Lettenmaier, D. P., Marshall, C. H., Entin, J. K., Pan, M., Shi, W., Koren, V., Meng, J., Ramsay, B. H., and Bailey, A. A.: The multi-institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system, J. Geophys. Res., 109, D07S90, doi:10.1029/2003JD003823, 2004.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models Part I — A discussion of principles, J. Hydrol., 10, 282–290, 1970.

Nandagiri, L.: Calibrating hydrological models in unguaged basins: possible use of areal evapotranspiration instead of streamflows, Predictions In Ungauged Basins: PUB Kick-off (Proceedings of the Kick-off meeting held in Brasilia, 20–22 November 2002) IAHS, 2007.

Nishida, K., Nemani, R. R., Running, S. W., and Glassy, J. M.: An operational remote sensing algorithm of land surface evaporation, J. Geophys. Res., 108, 4270, doi:10.1029/2002JD002062, 2003.

Oki, T., Musiake, K., Matsuyama, H., and Masuda, K.: Global atmospheric water balance and runoff from large river basins, Hydrol. Process., 9, 655–678, doi:10.1002/hyp.3360090513, 1995.

Pan, M., Sahoo, A. K., Troy, T. J., Vinukollu, R., Sheffield, J., and Wood, E. F.: Multi-source estimation of long-term Terrestrial Water Budget for major global river basins, J. Climate, 25, 3191–3206, doi:10.1175/JCLI-D-11-00300.1, 2012.

Pinker, R. T. and Laszlo, I.: Modeling surface solar irradiance for satellite applications on a global scale, J. Appl. Meteorol., 31, 194–211, 1992.

Rasmusson, E. M.: ATmospheric water vapor transport and the water balance of North America: Part I. Characteristics of the water vapor flux field, Mon. Weather Rev., 95, 403–426, doi:10.1175/1520-0493(1967)095<0403:AWVTAT>2.3.CO;2, 1967.

Rasmussen, E. M.: Atmospheric water vapor transport and the water balance of North America, 2, Large-scale water balance investigations, Mon. Weather Rev., 96, 720–734, 1968.

Ropelewski, C. F. and Yarosh, E. S.: The observed mean annual cycle of moisture budgets over the central United States (1973–92), J. Climate, 11 , 2180–2190, 1998.

Rosen, R. D. and Omolayo, A. S.: Exchange of water vapor between land and ocean in the Northern Hemisphere, J. Geophys Res., 86, 12147–12152, 1981.

Ruane, A. C.: NARR's atmospheric water cycle components — Part II: Summertime mean and diurnal interactions, J. Hydrometeorol., 11, 1220–1233, doi:10.1175/2010JHM1279.1, 2010.

Schaake, J., Cong S., and Duan, Q.: The US MOPEX Data Set, IAHS Red Book #307, 2006.

Schaefli, B. and Gupta, H. V.: Do Nash values have value?, Hydrol. Process., 21, 2075–2080, 2007.

Shukla, S, Steinemann, A. C., and Lettenmaier, D. P.: Drought Monitoring for Washington State: Indicators and Applications, J. Hydrometeor , 12, 66–83, doi:10.1175/2010JHM1307.1, 2011.

Son, K., and Sivapalan, M.: Improving model structure and reducing parameter uncertainty in conceptual water balance models through the use of auxiliary data, Water Resour. Res., 43, WO1415, doi:10.1029/2006WR005032, 2007.

Starr, V. P., Peixoto, J. P., and Crisi, H. R.: Hemispheric water balance for the IGY, Tellus, 17, 463–472, 1965.

Syed, T. H., Famiglietti, J. S., Chen, J., Rodell, M., Seneviratne, S. I., Viterbo, P., and Wilson, C. R.: Total basin discharge for the Amazon and Mississippi river basins from GRACE and a land-atmosphere water balance, Geophys. Res. Lett., 32, L24404, doi:10.1029/2005GL024851, 2005.

Tang, Q., Peterson, S., Cuenca, R. H., Hagimoto, Y., and Lettenmaier, D. P.: Satellite-based near-real-time estimation of irrigated crop water consumption, J. Geophys. Res. 114, D05114, doi:10.1029/2008JD010854, 2009.

Vano, J., Das, T., and Lettenmaier, D. P.: Hydrologic sensitivities of Colorado River runoff to changes in precipitation and temperature, J. Hydrometeorol., 13, 932–949, doi:10.1175/JHM-D-11-069.1, 2012.

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

Werth, S. and Güntner, A.: Calibration analysis for water storage variability of the global hydrological model WGHM, Hydrol. Earth Syst. Sci., 14, 59–78, doi:10.5194/hess-14-59-2010, 2010.

Wei, H., Xia, Y., Mitchell, K. E., and Ek, M. B.: Improvement of the Noah land surface model for warm season processes: evaluation of water and energy flux simulation. Hydrol. Process., doi:10.1002/hyp.9214, in press, 2012.

Xia, Y., Mitchell, K., Ek, M., Cosgrove, B., Sheffield, J., Luo, L., Alonge, C., Wei, H., Meng, J., Livneh, B., Duan, Q., and Lohmann, D.: Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System Project Phase 2 (NLDAS-2), Part 2: Validation of Model-simulated streamflow, J. Geophys. Res., 117, D03110, doi:10.1029/2011JD016051, 2012.

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

Yeh, P. J. F. and Famiglietti, J. S.: Regional terrestrial water storage change and evapotranspiration from terrestrial and atmospheric water balance computations, J. Geophys. Res., 113, D09108, doi:10.1029/2007JD009045, 2008.

Yeh, P. J. F., Irizarry, M., and Eltahir, E. A. B.: Hydroclimatology of Illinois: A comparison of monthly evaporation estimates based on atmospheric water balance and soil water balance, J. Geophys. Res., 103, 19823–19837, 1998.

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