Benchmarking of a Physically Based Hydrologic Model

ANDREW J. NEWMAN, NAOKI MIZUKAMI, MARTYN P. CLARK, AND ANDREW W. WOOD

National Center for Atmospheric Research,[a] Boulder, Colorado

BART NIJSSEN

Department of Civil and Environmental Engineering, University of Washington, Seattle, Washington

GREY NEARING

National Center for Atmospheric Research,[a] Boulder, Colorado, and NASA Goddard Space Flight Center, Greenbelt, Maryland

(Manuscript received 6 December 2016, in final form 5 May 2017)

ABSTRACT

The concepts of model benchmarking, model agility, and large-sample hydrology are becoming more prevalent in hydrologic and land surface modeling. As modeling systems become more sophisticated, these concepts have the ability to help improve modeling capabilities and understanding. In this paper, their utility is demonstrated with an application of the physically based Variable Infiltration Capacity model (VIC). The authors implement VIC for a sample of 531 basins across the contiguous United States, incrementally increase model agility, and perform comparisons to a benchmark. The use of a large-sample set allows for statistically robust comparisons and subcategorization across hydroclimate conditions. Our benchmark is a calibrated, time-stepping, conceptual hydrologic model. This model is constrained by physical relationships such as the water balance, and it complements purely statistical benchmarks due to the increased physical realism and permits physically motivated benchmarking using metrics that relate one variable to another (e.g., runoff ratio). The authors find that increasing model agility along the parameter dimension, as measured by the number of model parameters available for calibration, does increase model performance for calibration and validation periods relative to less agile implementations. However, as agility increases, transferability decreases, even for a complex model such as VIC. The benchmark outperforms VIC in even the most agile case when evaluated across the entire basin set. However, VIC meets or exceeds benchmark performance in basins with high runoff ratios (greater than ~0.8), highlighting the ability of large-sample comparative hydrology to identify hydroclimatic performance variations.

1. Introduction

Hydrologic and land models typically evolve through an iterative process of model refinement, evaluation, and diagnosis. This process necessarily requires testing these models against data. Traditionally, we compare the performance of a set of models at reproducing the observations; for example, via likelihood ratios or Bayesian model selection methods, or less rigorously by calculating error statistics across several competing models (e.g., Henderson-Sellers et al. 1993, 1995; Duan et al. 2006; Schlosser et al. 2000; Koster et al. 2004).

Recently, the hydrology and land modeling communities have made greater use of the scientific method in model evaluation to reject a model as inappropriate given the available observational data (Peters et al. 2003; Vaché and McDonnell 2006; Abramowitz et al. 2008; Best et al. 2015). The hydrology and land model evaluation communities refer to this process as "model benchmarking" (e.g., Abramowitz 2005; Abramowitz et al. 2008; Luo et al. 2012; Best et al. 2015; Nearing et al. 2016). A benchmark consists of the a priori data, the model, and derived expectation of performance against which we test other models. Benchmarking differs from other types of model evaluation because we use the a priori model to define

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JHM-D-16-0284.s1.

[a] The National Center for Atmospheric Research is sponsored by the National Science Foundation.

Corresponding author: Andrew Newman, anewman@ucar.edu

DOI: 10.1175/JHM-D-16-0284.1

our expectations of performance, yet it is similar to traditional model evaluation in that we are still comparing models across metrics based on observations.

This paper uses the benchmarking perspective to evaluate a commonly used hydrologic model, the Variable Infiltration Capacity model (VIC; Liang et al. 1994). The choice of the particular model is not especially important, outside of our requirement that it be physically motivated. We are interested in understanding the extent to which a physically motivated model uses available information and to identify opportunities for model improvement. We evaluate if VIC is sufficiently agile to reproduce the observations, as recent work has shown that physically motivated hydrologic models are too inflexible and thus limit their performance capabilities (Mendoza et al. 2015; Cuntz et al. 2016). We do this in two ways:

1) We systematically increase the agility of VIC by providing flexibility in model parameters that are typically neglected in VIC calibration efforts. The ability of a model to simulate a variety of different systems is referred to as ''model agility'' (Mendoza et al. 2015).
2) We conduct analyses across the contiguous United States (CONUS) covering a broad range of hydroclimatic regimes using many watersheds (Newman et al. 2014, 2015). Use of many basins allows for robust statistical comparisons and identification of systematic model performance issues (e.g., Duan et al. 2006; Newman et al. 2015). Such effort in comparative hydrology using many basins (e.g., more than 30) is coined ''large-sample hydrology'' (Gupta et al. 2014).

For our benchmark, we calibrated a parsimonious bucket-style hydrologic model across many watersheds. This model has fewer constraints in the form of physical process descriptions than VIC, and therefore presumably has more flexibility in reproducing observed behavior through parameter estimation. Bucket-style hydrologic models are a useful benchmark because they include some physical understanding of hydrologic systems, and, unlike many statistical benchmarks, bucket-style models include balance constraints. Bucket-style hydrologic models are also typically low dimensional (in terms of model states, parameters, and input data), and there has been much research into the calibration of this type of model (e.g., Duan et al. 1992; Yapo et al. 1998). Bucket-style models therefore provide a useful benchmark because they are easily applied and widely understood and can provide a practical prior that includes the basic hypothesis of continuity (water balance).

The central question that we are asking here is, ''Do the physical parameterizations in the more complex model improve our ability to simulate a variety of watersheds under a variety of conditions?'' We ask this question to evaluate the ability of large-sample hydrology and benchmarking to provide useful insights on model performance.

The remainder of the paper is organized as follows. In the next section we briefly describe the input dataset, basins, and the a priori conceptual hydrologic model, and then we describe VIC and experimental design. We then step through an iterative calibration and evaluation of VIC, examining multiple calibration configurations affording increasing model agility, each of which is compared to the benchmark using several illustrative metrics.

2. Watershed dataset and benchmark

It is advantageous to develop benchmarks for as wide a range of hydroclimatic conditions as possible. Large-sample hydrologic studies enable benchmarking of models across a wide range of hydroclimatic conditions, creating opportunities for more robust hypothesis testing and identifying satisfactory or unsatisfactory model systems or components. We use the term ''large sample'' to indicate a collection of basins greater than 30 that would ideally span as large a range of basin characteristics as possible. These large samples of basins allow for statistically meaningful statements to be made using comparative hydrology (Gupta et al. 2014). Modelers are increasingly able to perform such studies because growing computational resources have enabled so-called large-sample and large-domain hydrologic modeling studies and datasets. These began with macroscale regional and continental hydrologic modeling efforts of the early 2000s (e.g., Maurer et al. 2002; Lohman et al. 2004), and more recently have yielded several large collections of individual watersheds (e.g., Gupta et al. 2014, and references therein; Chaney et al. 2015; Rakovec et al. 2016).

In this study, we achieve broad and varied spatial coverage using a large-sample, basin-scale hydrometeorology dataset developed by Newman et al. (2014, 2015). This dataset comprises daily time step, basin-mean forcing data, calibrated conceptual watershed models and their daily model output, and observed streamflow observations for 671 basins across the CONUS. We use the dataset to define the basins used, their spatial extent, meteorological forcing data, and observed streamflow data for calibration and validation for all experiments. Figure 1

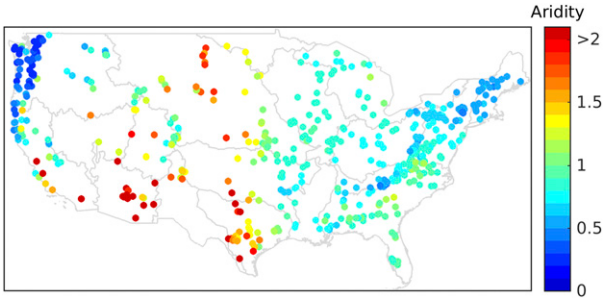

FIG. 1. The distribution of the 531 basins used in this study from Newman et al. (2014, 2015) and their corresponding estimated observed aridity indices.

highlights the basin locations and their aridity index. More details regarding specific usage of components of this dataset are given in the experimental design (section 3).

The Snow-17 temperature index snow accumulation and melt model (Anderson 1973), the Sacramento Soil Moisture Accounting (SAC-SMA) model (Burnash et al. 1973; Burnash 1995), and a unit hydrograph routing scheme are combined into a complete conceptual hydrologic modeling system, hereafter referred to as the River Forecast System (RFS). The RFS modeling system is our a priori model. We derived values of several illustrative metrics [daily Nash–Sutcliffe efficiency (NSE), volume bias, seasonality bias, and flow variability; see section 4 for details] from the calibrated RFS to define our benchmark performance metrics for adequate VIC performance.

We argue that selecting a calibrated conceptual hydrologic model as a benchmark is a useful practical application of the concepts in Nearing and Gupta (2015), that a benchmark should provide proper context for testing a model. In our case, the system is the RFS, but other simple conceptual modeling schemes may equally serve this purpose. This type of model is suitable for use in benchmarking because 1) a robust global optimization routine can extract a large amount of information, meaning that the calibrated RFS is functionally similar to a purely statistical reference model (e.g., Abramowitz 2012; Best et al. 2015); 2) unlike many statistical models, the RFS closes the water balance [$P = \mathrm{ET} + Q + \Delta S$; for precipitation P, evapotranspiration (ET), discharge Q, and storage S] and preserves important interactions between state variables and other observed features of hydrologic systems; and 3) the RFS has been and is still in wide use across the hydrologic community, providing a familiar benchmark for the hydrologic sciences community. The RFS is much less complex and computationally intensive than VIC, even in the greatly simplified VIC configuration we use here. With water balance and state interactions within the a priori model, we are also able to perform physically motivated benchmarking using metrics that relate one variable to another (e.g., runoff ratio).

3. Experimental design

Our experiment compares four implementations of VIC, each with different parameter flexibility and corresponding agility, versus the RFS-derived metrics, across the collection of watersheds. The objectives are to determine the impact of parameter-related model agility on VIC's effectiveness in using information as measured by common metrics and to identify deficits in performance.

We adopt a split-sample calibration (water years 2000–08) and validation (water years 1990–99) approach using 10 years of continuous model warm up prior to the calibration and validation phases. We subset the complete Newman et al. (2014) basin list to remove basins with nontrivial area discrepancies between the geospatial fabric developed by the USGS Modeling of Watersheds group (Viger and Bock 2014) and the USGS Geospatial Attributes of Gages for Evaluating Streamflow, version 2 (GAGES-II; Falcone 2011), basin boundaries (Bock et al. 2016) as well as basins larger than 2000 km^2. Table S1.1 (see the online supplemental material) contains a list of all 531 basins used. Additionally, we hold constant the snow correction factor (SCF) and rain/snow threshold temperature from the Snow-17 model (at 1 and 273.15 K, respectively), leaving 18 calibrated parameters in the RFS in this configuration (Table S1.2). Fixing SCF keeps the forcing input data consistent between VIC and RFS (i.e., we impose the constraint that the RFS cannot scale its precipitation input, as occurs in many RFS applications).

Both the RFS and VIC parameters are optimized via the shuffled complex evolution (SCE) algorithm (Duan et al. 1992, 1993) to minimize the daily root-mean-square error (RMSE) between simulated and observed flows. For VIC, calibration uses the multiscale parameter regionalization (MPR) estimation method described in Samaniego et al. (2010) for key VIC soil parameters (Table 1). Vegetation and elevation band parameter fields were created by merging default gridded VIC vegetation parameters within four elevation bands per basin. Elevation bands within the basin allow subbasin heterogeneity in precipitation and temperature while keeping the basin-mean values unchanged. This is considered a model

TABLE 1. VIC and routing model parameters used in different calibration cases. In the "Estimation method" column, VIC parameters estimated via MPR are denoted with MPR, parameters estimated via application of multiplier of basin-mean value are denoted with MULT, and direct parameter calibration is denoted with DIR.

Case	VIC parameter name	Description	Estimation method	Data source
1	Unit hydrograph scale	Scale of unit hydrograph gamma delay routing	DIR	—
	Unit hydrograph shape	Shape of unit hydrograph gamma delay routing	DIR	—
2	b_{infilt}	VIC infiltration parameter	MPR	STATSGO[b]
	$D1$	First baseflow parameter[a]	MPR	STATSGO[b]
	$D2$	Second baseflow parameter[a]	MPR	STATSGO[b]
	$D3$	Third baseflow parameter[a]	MPR	STATSGO[b]
	Depth soil layer 2	Depth of second VIC soil layer	MPR	—
	Depth soil layer 3	Depth of third VIC soil layer	MPR	—
	Total soil depth	Depth of total VIC soil column	MPR	STATSGO[b]
3	K_s	Saturated hydrologic conductivity	MPR	STATSGO[b]
	Bulk density	Bulk density of soil (used in VIC estimation of porosity)	MPR	STATSGO[b]
4	C_{min}	Min stomatal resistance	MULT	Maurer et al. (2002)
	LAI	Leaf area index	MULT	Maurer et al. (2002)

[a] From Nijssen et al. (2001).
[b] From State Soil Geographic Database (Miller and White 1998).

structural feature of VIC as it is a key component of VIC design. Default VIC parameters were taken from the Bureau of Reclamation/U.S. Army Corps of Engineers VIC CMIP5 climate projection simulations (Brekke et al. 2014; Wood and Mizukami 2014). We use basin-mean forcing derived from the 1/8° Maurer et al. (2002) forcing dataset compiled in Newman et al. (2014) for all experiments, allowing any performance variations to be attributable to model structure and physics.

VIC, version 4.1.2h, was calibrated four times, each with an increasing number of model parameters, to demonstrate the impact of the increasing model agility afforded by the additional parameter flexibility. The first case (Table 1, case 1) holds fixed the default semicalibrated VIC parameters (inherited from prior studies, many of which used earlier VIC versions) and allows calibration of only the routing model via the parameters for the shape and scale of the unit hydrograph, a gamma function. This model configuration is equivalent to saying our default physically motivated model (and its parameters) encode our knowledge of the hydrologic systems in question. Case 2 allows calibration of case 1 parameters and nearly all of the commonly calibrated VIC soil parameters (e.g., Nijssen et al. 2001; Demaria et al. 2007; Shi et al. 2008; Troy et al. 2008; Elsner et al. 2014; Oubeidillah et al. 2014). Case 3 includes case 1 and 2 parameters as well as soil porosity and saturated hydraulic conductivity. Finally, case 4 includes all previous case parameters as well as minimum stomatal resistance and monthly varying LAI, which are vegetation parameters that control evapotranspiration (for a total of 13 parameters).

These four cases were developed to explore VIC performance and agility through the lens of historical VIC (and other complex model) calibration efforts. For context, in the past, the additional case 4 parameters were not calibrated because they were considered more directly observable, hence "constrained," than the soil parameters. However, the default values are from specific (and limited) observation periods and locations and contain large uncertainties. Case 3 soil parameters were not calibrated because they were considered mostly duplicative of the effects of the case 2 parameters. Regardless of the reasoning behind the selection of calibration parameters, each of our successive cases enlarges the envelope of potential model behavior (its agility) given fixed forcing and a prescribed calibration objective function. Adding case 4 parameters allows substantially greater control over the water balance by offering a direct influence on evapotranspiration rates. Without case 4 parameters, ET is most commonly influenced in VIC indirectly by altering soil depths, which influences the amount and timing of soil water availability for ET, and vertical drainage rates. However, soil depths affect other characteristics of the rainfall runoff responses as well, making them a more complex model control than vegetation parameters. In many LSMs, in contrast to the practice surrounding VIC, soil depths are fixed, and vegetation parameters controlling ET are more often altered (e.g., Noah-MP; Niu et al. 2011). Finally, past practice has been to calibrate fewer parameters in higher-order complexity models such as VIC than in conceptual models for two main reasons. Model parameters were attributed physical meaning in the former class of models and their run times are often many times slower.

4. Results

Here we discuss the impacts of VIC parameter agility on model performance using four metrics derived from daily observed streamflow, NSE (Nash and Sutcliffe 1970), volume bias, relative flow variability (using the standard deviation ratio), and a flow seasonality metric compared to each other using case 1 as the base model implementation (section 4a). This evaluation of VIC flexibility is followed by a comparison of VIC with the benchmark based on the entire basin set [e.g., probability density functions (PDFs); section 4b(1)] along with spatial comparisons [section 4b(2)]. Finally, we highlight basins for which VIC has significant performance deviations using the NSE metric [section 4b(3)]. Note that because SCE is a single calibration approach and we calibrated to daily streamflow observations, these results should not be generalized to other variables (e.g., latent heat flux), as they are not well constrained by this calibration approach (e.g., Mendoza et al. 2015; Rakovec et al. 2016).

a. *Model agility*

Figure 2 shows the cumulative density function (CDF) of the NSE scores for each of the VIC calibrations cases for the 531 study watersheds for both the calibration (dashed) and validation (solid) periods. As expected, calibrating only the routing model parameters in case 1 yields inferior performance. However, one interesting feature of case 1 is increased validation period performance. Anecdotally, we believe this is at least partially because some of the semicalibrated case 1 VIC parameters were developed in the 1990s and early 2000s and used the 1990s and prior decades as their calibration period. For this study, the 2000s are the calibration period and the 1990s are the validation period, which implies that we are likely using at least part of the calibration period for the default VIC parameters in at least some basins for our validation period when validating case 1.

Increasing model agility by increasing the number of calibrated parameters progressively improves model performance in both the calibration and validation periods (Fig. 2a). The colored lines are incrementally farther to the right in the plots for the calibration and validation phases when comparing more agile to less agile VIC configurations (e.g., case 4 to case 1). Each additional parameter provides less improvement (Fig. 2b), which may be a result of insensitive or poorly identifiable parameters being selected for optimization (e.g., Demaria et al. 2007), or there is less information available for extraction as the model agility increases.

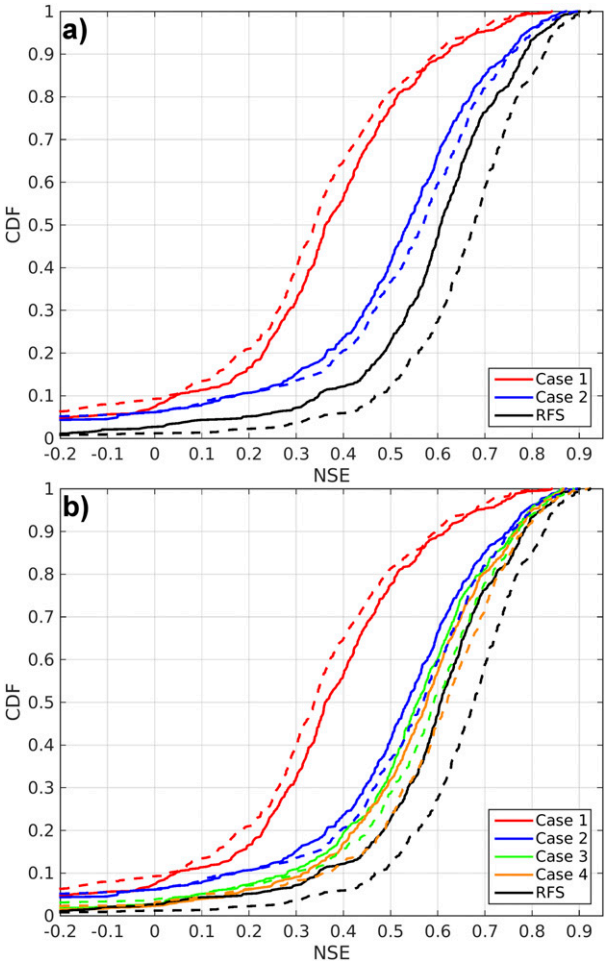

FIG. 2. CDFs of NSE for the VIC (a) cases 1 and 2 and (b) all four cases. The RFS benchmark NSE (black lines) is included in both panels. Solid (dashed) lines indicate performance in validation (calibration) time periods.

However, proper exploration of the order of parameters in the calibration cases, their value in calibration, and which parameters are potentially identifiable requires formal sensitivity analysis and many variations in the calibration scenarios in Table 1. This is an area of active research and will be explored in the future. It is highly likely that changing the order of the cases and/or VIC parameters in the calibrations will lead to slightly different results and the appearance of different higher-importance parameters. However, the primary subject of the paper is to explore ability of large-sample hydrology and a conceptual model benchmark to provide statistically robust statements on model performance through the lens of model agility in physically motivated hydrologic models. The order of calibration is not important to explore the general impact of increased model agility. Thus, we will avoid making any additional statements on the significance of performance

differences between cases 2 and 4 to avoid implication that one VIC parameter (or parameter set) is more valuable than another.

The benefits of increased model agility are evident not only in the CDFs of NSE scores (Fig. 2), but also across other hydrograph-related metrics. Figure 3a shows the PDFs of normalized total flow biases for the four VIC cases, and Fig. 3b displays the relative flow variability using the standard deviation ratio: σ(model)/σ(observations) (e.g., Gupta et al. 2009). Improvement in both metrics coincides with increased agility as evidenced by narrower PDFs as well as median values that are closer to zero or one for flow bias and the variability ratio, respectively.

Table 2 summarizes VIC performance for each case during the validation period. We assess whether the marginal parameter additions of each successive case provide statistically significant improvements in performance at the 90% confidence level by estimating four median-based statistics: 1) the median NSE, 2) median absolute flow volume bias, 3) median absolute flow seasonality bias, and 4) median absolute relative flow variability ratio (deviations from 1, smaller numbers denote better model performance), with 90% confidence intervals estimated via bootstrapping with 1000 samples. For NSE only, we also use the one-sided Kolmogorov–Smirnov (KS) test to compare the VIC NSE CDFs with each other. A one-sided KS test will identify CDFs for which the shift toward higher values is statistically significant (again at the 90% level) and provides a distribution-based statistical test rather than a simple comparison of changes in median values.

Increasing VIC agility improves the median performance across the basin set. All cases that include VIC parameters in the calibration (cases 2–4) are significantly better than case 1 for all median metrics except case 2 for seasonality bias. Additionally, case 4 is significantly better than case 2 for NSE, implying that further increases in model agility are resulting in clear performance increases. Minimal changes in volume bias and the absolute relative flow variability ratio are expected among cases 2–4 because the objective function (daily RMSE) minimally constrains bias, and the optimal RMSE calibration results in an underprediction of the variance (Gupta et al. 2009). Thus, once VIC is agile enough to obtain a reasonable solution, further improvement in absolute relative flow variability is limited. Finally, the one-sided KS test indicates that cases 2–4 are significantly better than case 1 across the entire NSE distributions.

b. *VIC benchmarking*

Section 4a clearly shows that increasing the agility of VIC provides significant improvements across a large basin set, even above the typical soil calibration (case 2). However, model improvements are only demonstrated relative to case 1. It is useful to assess how well these VIC cases perform relative to a model in which fewer physical relationships are encoded. Comparisons using observed streamflow from the entire large-sample basin set give general model performance characteristics, while spatial and hydroclimatic comparisons more robustly identify characteristics of model deficiencies. With this in mind, we compare VIC performance against the RFS benchmark.

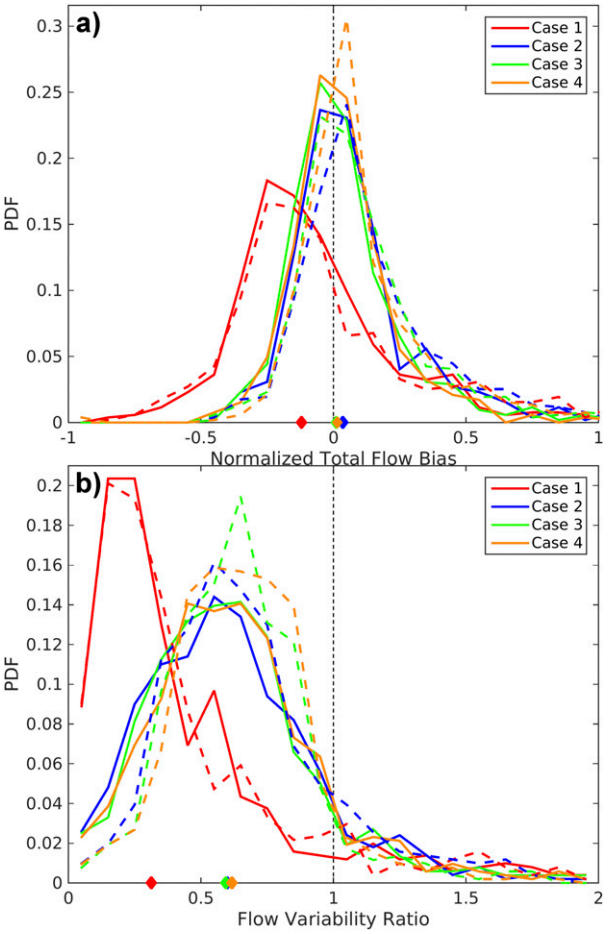

FIG. 3. PDFs of (a) normalized flow volume bias and (b) flow variability ratio. The diamonds along the x axis indicate the median of each PDF. Solid (dashed) lines indicate performance in validation (calibration) time periods.

1) OVERVIEW COMPARISONS

Case 2 yields better results than the base case, but its performance is clearly worse than the benchmark (Fig. 2a). Even with the optimization of routing, soil, and vegetation parameters (case 4), VIC is still inferior to the benchmark across the basin set as a whole (Fig. 2b).

TABLE 2. Validation period summary metrics from the four VIC calibration cases and the RFS a priori (configuration). Ranges given are the 90% confidence intervals estimated from bootstrapping with replacement ($N = 1000$). Results are rounded to two significant digits for readability, while statistical tests are applied to full precision values.

Configuration	NSE (median)	Volume bias	Seasonality bias	Flow variability ratio
Case 1	0.35–0.39	0.20–0.23	0.43–0.52	0.70–0.74
Case 2	0.52–0.55	0.10–0.12	0.41–0.50	0.44–0.48
Case 3	0.54–0.57	0.09–0.11	0.34–0.41	0.43–0.47
Case 4	0.57–0.59	0.09–0.11	0.33–0.39	0.41–0.46
RFS	0.60–0.62	0.13–0.15	0.27–0.33	0.39–0.43

The last row of Table 2 includes the RFS summary statistics. VIC case 4 significantly underperforms the RFS across the basin set for median NSE and across the distribution of NSE using the KS test, with nonsignificant differences in median absolute seasonality and relative flow variability ratio. Case 4 has better performance for median absolute volume bias.

We note that there are larger NSE differences between the calibration and validation period as the agility of a model increases, suggesting more overfitting to noise as the agility increases. Figure 4 highlights the PDFs for the NSE differences (calibration minus validation) for the four VIC configurations and the RFS. The NSE differences for VIC cases 2–4 (90% confidence interval) are 0.010–0.022, 0.013–0.027, and 0.021–0.036, respectively. Although the VIC cases are not significantly different (excepting case 1), there is a clear trend of increasing NSE differences with increasing agility. Additionally, the median NSE difference for the RFS is 0.044–0.057, which is significantly larger than the VIC cases 2–4 differences. This result hints at a trade-off that increasing model agility may result in less robustness (i.e., overfitting): parameter flexibility is being used to fit the model to noise in the input/output data sample in the RFS benchmark, and even for complex models such as VIC in which a smaller fraction of uncertain model parameters are exposed to calibration.

2) SPATIAL COMPARISONS

Figure 5 shows the geographic variations in NSE for VIC case 4 and the RFS. VIC and the RFS performance varies in a spatially similar manner, with high NSE (>0.7) across the windward Pacific Northwest (PNW; upper-left portion of Fig. 5) and NSE > 0.5 across most of the eastern third of CONUS. Along the drier high plains (strip of blue just west of the central United States) and desert southwest (lower-left portion of Fig. 5), both perform worse. Note that the better performance of the RFS in Fig. 2 across all quantiles of the CDF does not imply that VIC is worse for all locations. The difference field between VIC case 4 and the RFS (Fig. 5c) is consistent with the CDFs in showing that VIC is generally underperforming (Fig. 2, Table 2). From Fig. 5c, the differences between VIC case 4 and the benchmark appear to be randomly distributed throughout CONUS, with some tendency for VIC to perform worse in basins across the western half of the domain outside of the windward basins along the PNW coast. Basins in the PNW with low VIC skill nearly all fall on the lee of the two mountain ranges in the region, which are much drier basins. The windward PNW coastline is extremely wet, while the lee slopes of the PNW and most of the rest of the western half of the United States are arid to semiarid.

3) HYDROCLIMATIC COMPARISONS

Both models perform better as runoff ratio increases (not shown). Figure 6 shows pairwise differences of VIC case 4 minus the RFS NSE as a function of the estimated observed runoff ratio (using the forcing precipitation). It shows a tendency for VIC to perform worse relative to the RFS for basins with a low runoff ratio than for basins with higher runoff ratios. Binning

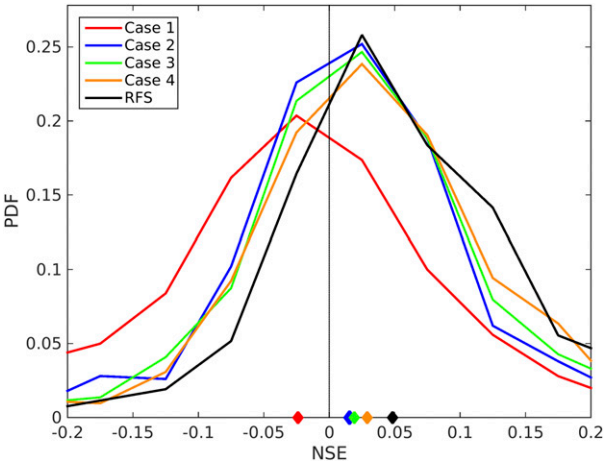

FIG. 4. PDFs of NSE differences (calibration minus validation) from the four VIC cases and the RFS. The diamonds along the x axis indicate the median of each PDF.

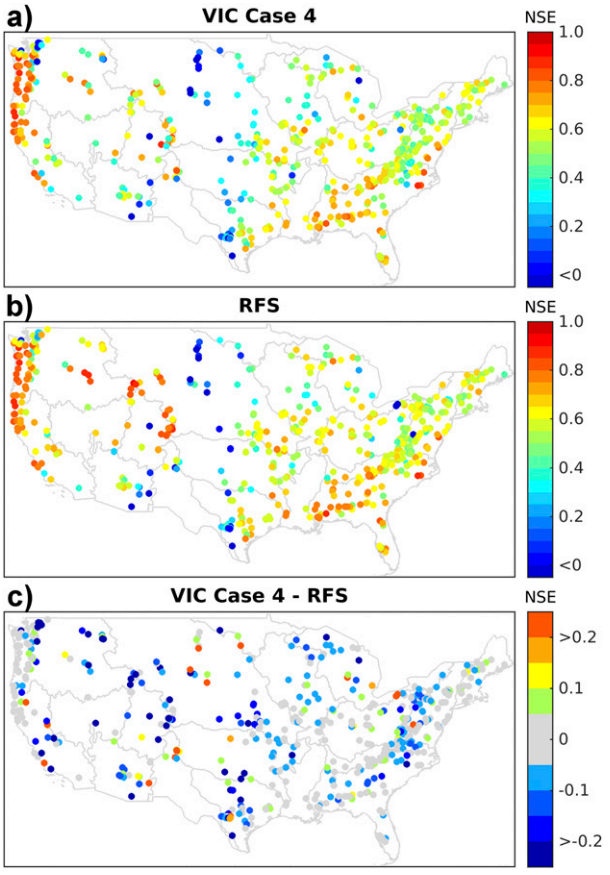

FIG. 5. (a) VIC case 4 NSE, (b) RFS NSE, (c) VIC case 4 minus RFS NSE difference field. Gray lines denote USGS hydrologic unit code level 2 boundaries.

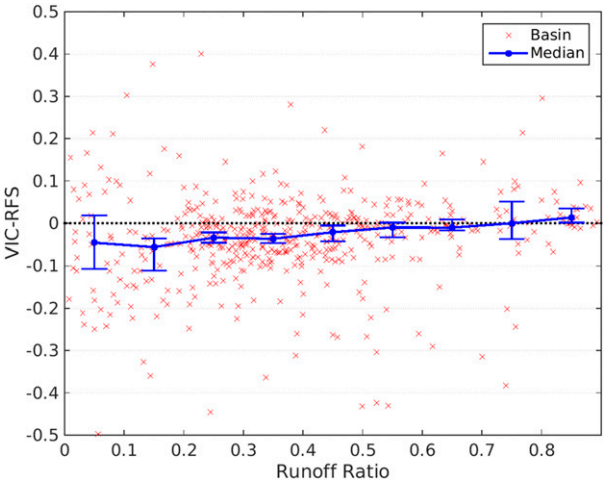

FIG. 6. NSE difference (case 4 minus RFS) for all 531 basins (red crosses) and median difference with 90% confidence intervals derived from bootstrapping (blue dots are median values, confidence interval given by vertical blue lines) for bins (width of 0.1) of basins by runoff ratio.

the basins by runoff ratio (bin sizes of 0.1 runoff ratio) and computing median NSE differences and 90% confidence intervals for all basins in a given bin using bootstrapping helps illustrate this tendency. There is a significant difference in favor of the RFS for nearly all bins with runoff ratios less than 0.7, and there is a clear relationship between VIC performance and runoff ratio. Finally, VIC outperforms the RFS for basins with runoff ratios greater than 0.8 at the 90% level, although the sample size of these basins is relatively small (26 basins).

Most surface hydrological models perform worse for runoff generation in arid basins, particularly in daily time step and watershed-scale configurations. The models may miss processes such as surface–groundwater interactions and channel losses, but more importantly, since runoff forms only a small component of the water balance, small relative errors in the simulation of evapotranspiration result in large relative errors in the simulation of runoff. Both of the models used here perform worse in more arid basins in general (e.g., Nijssen et al. 1997; Anderson 2002; Demaria et al. 2007; Newman et al. 2015). However, VIC performing worse relative to the RFS across arid basins highlights specific research avenues for VIC in arid catchments discussed in previous literature (e.g., Abdulla and Lettenmaier 1997; Nijssen et al. 1997; Demaria et al. 2007).

Figure 6 shows outliers in both directions. An outlier can be due to hydrologic model deficiency, errors in observation data, errors in model implementation (human coding errors), or other factors. This emphasizes the need to use large-sample studies; single-basin (or small sample) studies can be drastically impacted by outliers and may have no ability to identify them, especially when they are conducted without a benchmark. As an example of a human coding error, an earlier implementation of the calibration routine was incorrectly updating soil depth, resulting in reduced performance for case 2 relative to the benchmark, particularly for basins with calibrated NSE values less than roughly 0.3. Relative to case 1, only this erroneous case 2 still had improved performance across all basins. The reduced case 2 performance in the low NSE basin subset was only identified with the benchmark comparison.

5. Summary and discussion

The concepts of model benchmarking, model agility, and large-sample hydrology have received increasing attention in the literature recently (Abramowitz 2012; Gupta et al. 2014; Newman et al. 2015; Mendoza et al. 2015; Best et al. 2015; Nearing and Gupta 2015). As the

hydrologic modeling community's questions and modeling systems become more intricate, these concepts should be used to help improve our modeling capabilities. Before further discussion, we summarize key messages from this study: 1) increasing model agility is key to increase physically motivated model performance across a wide range of basin hydroclimatic conditions in a significant manner; 2) increased model agility likely results in decreased model transferability; 3) model benchmarking using large-sample comparative hydrology allows for many statistically robust comparisons; 4) a conceptual model benchmark allows us to ask if our physical parameterizations in a more complex model actually improve our ability to simulate a variety of watersheds under a variety of different conditions; and 5) as an example for VIC, performance increases as runoff ratio increases and actually surpasses the conceptual model benchmark.

The ability of a model to simulate a variety of different hydroclimate regimes and phenomena is referred to as model agility (Mendoza et al. 2015; Cuntz et al. 2016). This might include capabilities to modify the spatial structure, process parameterizations, and parameter values (Clark et al. 2011, 2015). A lack of model agility, as seen in many existing physically based hydrologic models, hinders model evaluation and model performance (Mendoza et al. 2015). It is clear that increased model flexibility or agility, meaning expanding the parameters available for calibration in this case, leads to increased model performance (Figs. 2, 3; Table 2). This is a largely expected result that parallels outcomes in statistical modeling, where increasing the degrees of freedom should increase model performance (Hawkins 2004). However, it may not be guaranteed in complex physical system modeling, where parameter optimization routines may not find the objective function global minimum because of the ill-posed nature of these optimization problems (e.g., Duan et al. 1992; McLaughlin and Townley 1996; Yapo et al. 1998; Vrugt et al. 2003; Arsenault et al. 2014). An interesting result from these agility experiments is that as model agility increases, the stability of the model performance appears to decrease [section 4b(1); Fig. 4]. This trade-off will need to be considered in regional model implementation or parameter estimation activities. These results illustrate the need to understand exactly how different physical hypotheses, approximations, parameterizations, and assumptions that are embedded in our models enable versus restrict our ability to apply these models to general classes of systems (e.g., continental-scale hydrology), and to understand whether the ability of our models to make accurate and reliable predictions really comes from our understanding of hydrological systems versus primarily from regression. What do our models really know about the governing processes of watersheds, and how useful is this information in making predictions?

Benchmarking is a critical component of model evaluation work. When combined with large sample sizes, benchmarking can support a statistically robust evaluation of model performance, facilitating the identification of implementation or numerical errors, and stronger conclusions, leading to greater insight into model deficiencies that may help guide future research and development effort. This is a strong argument for using more formalized benchmarking approaches discussed by Nearing and Gupta (2015). It is also advantageous to the community that we identify missteps or failures in our modeling endeavors along with the successes (e.g., Andréassian et al. 2010). For example, here VIC is generally less skillful relative to the RFS, but as runoff ratio increases, the skill of VIC increases, even slightly surpassing the RFS for very efficient basins (Fig. 6). This result agrees with past small-sample basin studies (e.g., Nijssen et al. 1997; Demaria et al. 2007) that show VIC performs worse in arid regions, but highlights that VIC performs worse relative to a benchmark that also has degraded performance in arid basins [section 4b(3); Fig. 6]. This could direct future studies to examine processes in VIC that generate runoff in arid basins. Additional studies could also examine the feasibility of compiling other observations across many of these basins to examine other important variables such as surface heat fluxes, soil moisture, etc., in this type of benchmarking framework.

Finally, we suggest that using a calibrated conceptual hydrologic model that is relatively low dimensional (in model parameters, states, and input data requirements), even if constrained by physical relationships, offers an easily applied and widely understood practical benchmark for the community. The inclusion of balance constraints and time lags into the a priori model allows us to ask whether the physical parameterizations in the more complex model improve our ability to simulate a variety of watersheds under a variety of different conditions. This is very similar to using the simple physically based Penman–Monteith or Manabe bucket models for estimating surface fluxes in Best et al. (2015). Here a higher level of information usage in the a priori model is obtained because it is calibrated using a robust global optimization routine. Yet, we likely do not estimate the maximum information extraction possible because the conceptual model encodes a priori structural and physical assumptions that may or may not optimally fit

the data. These assumptions likely result in errors and information loss (Gong et al. 2013; Nearing and Gupta 2015); thus, this benchmark does not represent the ideal data-driven maximum information extraction benchmark discussed in Nearing and Gupta (2015). Pursuit of such an ideal is a topic for future work.

Acknowledgments. The authors thank Olda Rakovec for his help with MPR. This work was funded by the Bureau of Reclamation, Cooperative Agreement R11AC80816, and the U.S. Army Corps of Engineers (USACE) Climate Preparedness and Resilience Program.

REFERENCES

Abdulla, F., and D. P. Lettenmaier, 1997: Development of regional parameter estimation equations for a macroscale hydrologic model. *J. Hydrol.*, **197**, 230–257, doi:10.1016/S0022-1694(96)03262-3.

Abramowitz, G., 2005: Towards a benchmark for land surface models. *Geophys. Res. Lett.*, **32**, L22702, doi:10.1029/2005GL024419.

——, 2012: Towards a public, standardized, diagnostic benchmarking system for land surface models. *Geosci. Model Dev.*, **5**, 819–827, doi:10.5194/gmd-5-819-2012.

——, R. Leuning, M. Clark, and A. J. Pitman, 2008: Evaluating the performance of land surface models. *J. Climate*, **21**, 5468–5481, doi:10.1175/2008JCLI2378.1.

Anderson, E. A., 1973: National Weather Service River Forecast System—Snow accumulation and ablation model. NOAA Tech. Memo. NWS HYDRO-17, 87 pp. [Available online at https://www.wcc.nrcs.usda.gov/ftpref/wntsc/H&H/snow/AndersonHYDRO17.pdf.]

——, 2002: Calibration of conceptual hydrologic models for use in river forecasting. NOAA Rep., 372 pp.

Andréassian, V., C. Perrin, E. Parent, and A. Bárdossy, 2010: The Court of Miracles of Hydrology: Can failure stories contribute to hydrological science? *Hydrol. Sci. J.*, **55**, 849–856, doi:10.1080/02626667.2010.506050.

Arsenault, R., A. Poulin, P. Côté, and F. Brissette, 2014: Comparison of stochastic optimization algorithms in hydrological model calibration. *J. Hydrol. Eng.*, **19**, 1374–1384, doi:10.1061/(ASCE)HE.1943-5584.0000938.

Best, M. J., and Coauthors, 2015: The plumbing of land surface models: Benchmarking model performance. *J. Hydrometeor.*, **16**, 1425–1442, doi:10.1175/JHM-D-14-0158.1.

Bock, A. R., L. E. Hay, G. J. McCabe, S. L. Markstrom, and R. D. Atkinson, 2016: Parameter regionalization of a monthly water balance model for the conterminous United States. *Hydrol. Earth Syst. Sci.*, **20**, 2861–2876, doi:10.5194/hess-20-2861-2016.

Brekke, L., A. Wood, and T. Pruitt, 2014: Downscaled CMIP3 and CMIP5 hydrology projections: Release of hydrology projections, comparison with preceding information, and summary of user needs. USBR Tech Memo., 110 pp. [Available online at http://gdo-dcp.ucllnl.org/downscaled_cmip_projections/techmemo/BCSD5HydrologyMemo.pdf.]

Burnash, R. J. C., 1995: The NWS River Forecast System—Catchment model. *Computer Models of Watershed Hydrology*, V. P. Singh, Water Resources Publications, 311–366.

——, R. L. Ferral, and R. A. McGuire, 1973: A generalized streamflow simulation system: Conceptual models for digital computers. Joint Federal and State River Forecast Center, U.S. National Weather Service, and California Department of Water Resources Tech. Rep., 204 pp.

Chaney, N. W., J. D. Herman, P. M. Reed, and E. F. Wood, 2015: Flood and drought hydrologic monitoring: The role of model parameter uncertainty. *Hydrol. Earth Syst. Sci.*, **19**, 3239–3251, doi:10.5194/hess-19-3239-2015.

Clark, M. P., D. Kavetski, and F. Fenicia, 2011: Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resour. Res.*, **47**, W09301, doi:10.1029/2010WR009827.

——, and Coauthors, 2015: A unified approach to hydrologic modeling: 1. Modeling concept. *Water Resour. Res.*, **51**, 2498–2514, doi:10.1002/2015WR017198.

Cuntz, M., J. Mai, L. Samaniego, M. Clark, V. Wulfmeyer, O. Branch, S. Attinger, and S. Thober, 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**, 10 676–10 700, doi:10.1002/2016JD025097.

Demaria, E. M., B. Nijssen, and T. Wagener, 2007: Monte Carlo sensitivity analysis of land surface parameters using the Variable Infiltration Capacity model. *J. Geophys. Res.*, **112**, D11113, doi:10.1029/2006JD007534.

Duan, Q., S. Sorooshian, and V. K. Gupta, 1992: Effective and efficient global optimization for conceptual rainfall–runoff models. *Water Resour. Res.*, **28**, 1015–1031, doi:10.1029/91WR02985.

——, V. K. Gupta, and S. Sorooshian, 1993: A shuffled complex evolution approach for effective and efficient optimization. *J. Optim. Theory Appl.*, **76**, 501–521, doi:10.1007/BF00939380.

——, and Coauthors, 2006: Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops. *J. Hydrol.*, **320**, 3–17, doi:10.1016/j.jhydrol.2005.07.031.

Elsner, M. M., S. Gangopadhyay, T. Pruitt, L. D. Brekke, N. Mizukami, and M. P. Clark, 2014: How does the choice of distributed meteorological data affect hydrologic model calibration and streamflow simulations? *J. Hydrometeor.*, **15**, 1384–1403, doi:10.1175/JHM-D-13-083.1.

Falcone, J. A., 2011: GAGES-II: Geospatial Attributes of Gages for Evaluating Streamflow. USGS, accessed 10 October 2013. [Available online at http://water.usgs.gov/GIS/metadata/usgswrd/XML/gagesII_Sept2011.xml.]

Gong, W., H. V. Gupta, D. Yang, K. Sricharan, and A. O. Hero III, 2013: Estimating epistemic and aleatory uncertainties during hydrologic modeling: An information theoretic approach. *Water Resour. Res.*, **49**, 2253–2273, doi:10.1002/wrcr.20161.

Gupta, H. V., H. Kling, K. K. Yilmaz, and G. F. Martinez-Barquero, 2009: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modeling. *J. Hydrol.*, **377**, 80–91, doi:10.1016/j.jhydrol.2009.08.003.

——, C. Perrin, G. Blöschl, A. Montanari, R. Kumar, M. Clark, and V. Andréassian, 2014: Large-sample hydrology: A need to balance depth with breadth. *Hydrol. Earth Syst. Sci.*, **18**, 463–477, doi:10.5194/hess-18-463-2014.

Hawkins, D. M., 2004: The problem of overfitting. *J. Chem. Inf. Comput. Sci.*, **44**, 1–12, doi:10.1021/ci0342472.

Henderson-Sellers, A., Z.-L. Yang, and R. E. Dickinson, 1993: The Project for Intercomparison of Land-Surface Parameteriza-

tion Schemes. *Bull. Amer. Meteor. Soc.*, **74**, 1335–1349, doi:10.1175/1520-0477(1993)074<1335:TPFIOL>2.0.CO;2.

——, A. J. Pitman, P. K. Love, P. Irannejad, and T. Chen, 1995: The Project for Intercomparison of Land Surface Parameterization Schemes (PILPS): Phases 2 and 3. *Bull. Amer. Meteor. Soc.*, **76**, 489–503, doi:10.1175/1520-0477(1995)076<0489:TPFIOL>2.0.CO;2.

Koster, R. D., and Coauthors, 2004: Regions of coupling between soil moisture and precipitation. *Science*, **305**, 1138–1140, doi:10.1126/science.1100217.

Liang, X., D. P. Lettenmaier, E. F. Wood, and S. J. Burges, 1994: A simple hydrologically based model of land-surface water and energy fluxes for general circulation models. *J. Geophys. Res.*, **99**, 14 415–14 428, doi:10.1029/94JD00483.

Lohmann, D., and Coauthors, 2004: Streamflow and water balance intercomparisons of four land surface models in the North American Land Data Assimilation System project. *J. Geophys. Res.*, **109**, D07S91, doi:10.1029/2003JD003517.

Luo, Y. Q., and Coauthors, 2012: A framework for benchmarking land models. *Biogeosciences*, **9**, 3857–3874, doi:10.5194/bg-9-3857-2012.

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

McLaughlin, D., and L. R. Townley, 1996: A reassessment of the groundwater inverse problem. *Water Resour. Res.*, **32**, 1131–1161, doi:10.1029/96WR00160.

Mendoza, P., M. Clark, M. Barlage, B. Rajagopalan, L. Samaniego, G. Abramowitz, and H. V. Gupta, 2015: Are we unnecessarily constraining the agility of complex process-based models? *Water Resour. Res.*, **51**, 716–728, doi:10.1002/2014WR015820.

Miller, D. A., and R. A. White, 1998: A conterminous United States multilayer soil characteristics dataset for regional climate and hydrology modeling. *Earth Interact.*, **2**, doi:10.1175/1087-3562(1998)002<0001:ACUSMS>2.3.CO;2.

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

Nearing, G. S., and H. V. Gupta, 2015: The quantity and quality of information in hydrologic models. *Water Resour. Res.*, **51**, 524–538, doi:10.1002/2014WR015895.

——, D. M. Mocko, C. D. Peters-Lidard, S. V. Kumar, and Y. Xia, 2016: Benchmarking NLDAS-2 soil moisture and evapotranspiration to separate uncertainty contributions. *J. Hydrometeor.*, **17**, 745–759, doi:10.1175/JHM-D-15-0063.1.

Newman, A. J., K. Sampson, M. P. Clark, A. Bock, R. J. Viger, and D. Blodgett, 2014: CAMELS: Large-sample hydrometeorological dataset. NCAR Computational and Information Systems Laboratory Research Data Archive, accessed 7 March 2016, doi:10.5065/D6MW2F4D.

——, and Coauthors, 2015: Development of a large-sample watershed-scale hydrometeorological dataset for the contiguous USA: Dataset characteristics and assessment of regional variability in hydrologic model performance. *Hydrol. Earth Syst. Sci.*, **19**, 209–223, doi:10.5194/hess-19-209-2015.

Nijssen, B., D. P. Lettenmaier, X. Liang, S. W. Wetzel, and E. F. Wood, 1997: Streamflow simulation for continental-scale river basins. *Water Resour. Res.*, **33**, 711–724, doi:10.1029/96WR03517.

——, G. M. O'Donnell, D. P. Lettenmaier, D. Lohmann, and E. F. Wood, 2001: Predicting the discharge of global rivers. *J. Climate*, **14**, 3307–3323, doi:10.1175/1520-0442(2001)014<3307:PTDOGR>2.0.CO;2.

Niu, G.-Y., and Coauthors, 2011: The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *J. Geophys. Res.*, **116**, D12109, doi:10.1029/2010JD015139.

Oubeidillah, A. A., S.-C. Kao, M. Ashfaq, B. S. Naz, and G. Tootle, 2014: A large-scale, high-resolution hydrological model parameter data set for climate change impact assessment for the conterminous US. *Hydrol. Earth Syst. Sci.*, **18**, 67–84, doi:10.5194/hess-18-67-2014.

Peters, N. E., J. Freer, and K. Beven, 2003: Modelling hydrologic responses in a small forested catchment (Panola Mountain, Georgia, USA): A comparison of the original and a new dynamic TOPMODEL. *Hydrol. Processes*, **17**, 345–362, doi:10.1002/hyp.1128.

Rakovec, and Coauthors, 2016: Multiscale and multivariate evaluation of water fluxes and states over European river basins. *J. Hydrometeor.*, **17**, 287–307, doi:10.1175/JHM-D-15-0054.1.

Samaniego, L., A. Bárdossy, and R. Kumar, 2010: Streamflow prediction in ungauged catchments using copula-based dissimilarity measures. *Water Resour. Res.*, **46**, W02506, doi:10.1029/2008WR007695.

Schlosser, C. A., and Coauthors, 2000: Simulations of a boreal grassland hydrology at Valdai, Russia: PILPS phase 2(d). *Mon. Wea. Rev.*, **128**, 301–321, doi:10.1175/1520-0493(2000)128<0301:SOABGH>2.0.CO;2.

Shi, X., A. W. Wood, and D. P. Lettenmaier, 2008: How essential is hydrologic model calibration to seasonal streamflow forecasting? *J. Hydrometeor.*, **9**, 1350–1363, doi:10.1175/2008JHM1001.1.

Troy, T. J., E. F. Wood, and J. Sheffield, 2008: An efficient calibration method for continental-scale land surface modeling. *Water Resour. Res.*, **44**, W09411, doi:10.1029/2007WR006513.

Vaché, K. B., and J. J. McDonnell, 2006: A process-based rejectionist framework for evaluating catchment runoff model structure. *Water Resour. Res.*, **42**, W02409, doi:10.1029/2005WR004247.

Viger, R. J., and A. Bock, 2014: GIS features of the geospatial fabric for national hydrologic modeling. U.S. Geological Survey, accessed 14 December 2015, doi:10.5066/F7542KMD.

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

Wood, A. W., and N. Mizukami, 2014: CMIP5 1/8th degree daily weather and VIC hydrology datasets for CONUS. NCAR Final Project Rep. to USACE Responses to Climate Change, Project W26HM423495778, 32 pp. [Available online at http://www.corpsclimate.us/docs/cmip5.hydrology.2014.final.report.pdf.]

Yapo, P. O., H. V. Gupta, and S. Sorooshian, 1998: Multi-objective global optimization for hydrologic models. *J. Hydrol.*, **204**, 83–97, doi:10.1016/S0022-1694(97)00107-8.