In Quest of Calibration Density and Consistency in Hydrologic Modeling: Distributed Parameter Calibration against Streamflow Characteristics

Yuan Yang1,2, Ming Pan2, Hylke E. Beck2, Colby K. Fisher2, R. Edward Beighley3, Shih-Chieh Kao4, Yang Hong1,5, and Eric F. Wood2

1State Key Laboratory of Hydroscience and Engineering, Department of Hydraulic Engineering, Tsinghua University, Beijing, China, 2Department of Civil and Environmental Engineering, Princeton University, Princeton, New Jersey, USA, 3Department of Civil and Environmental Engineering, Northeastern University, Boston, Massachusetts, USA, 4Oak Ridge National Laboratory, Environmental Sciences Division and Climate Change Science Institute, Oak Ridge, Tennessee, USA, 5School of Civil Engineering and Environmental Sciences, University of Oklahoma, Norman, Oklahoma, USA

Abstract

Conventional basin-by-basin approaches to calibrate hydrologic models are limited to gauged basins and typically result in spatially discontinuous parameter fields. Moreover, the consequent low calibration density in space falls seriously behind the need from present-day applications like high resolution river hydrodynamic modeling. In this study we calibrated three key parameters of the Variable Infiltration Capacity (VIC) model at every 1/8° grid-cell using machine learning-based maps of four streamflow characteristics for the conterminous United States (CONUS), with a total of 52,663 grid-cells. This new calibration approach, as an alternative to parameter regionalization, applied to ungauged regions too. A key difference made here is that we tried to regionalize physical variables (streamflow characteristics) instead of model parameters whose behavior may often be less well understood. The resulting parameter fields no longer presented any spatial discontinuities and the patterns corresponded well with climate characteristics, such as aridity and runoff ratio. The calibrated parameters were evaluated against observed streamflow from 704/648 (calibration/validation period) small-to-medium-sized catchments used to derive the streamflow characteristics, 3941/3809 (calibration/validation period) small-to-medium-sized catchments not used to derive the streamflow characteristics as well as five large basins. Comparisons indicated marked improvements in bias and Nash-Sutcliffe efficiency. Model performance was still poor in arid and semiarid regions, which is mostly due to both model structural and forcing deficiencies. Although the performance gain was limited by the relative small number of parameters to calibrate, the study and results here served as a proof-of-concept for a new promising approach for fine-scale hydrologic model calibrations.

1. Introduction

Land surface models (LSMs) and distributed hydrologic models have been the major tool to study various hydrologic variables of interest at large scales from basin to globe and provide many critical monitoring/forecast services and risk analysis in water resources management, hazard mitigation (e.g., flood and drought), and climate adaptation (Asadieh & Krakauer, 2017; Hundecha et al., 2016; Madsen et al., 2014). Some models estimate river streamflow directly and others provide grid-cell or catchment level runoff, which is later fed to river hydrodynamic models for streamflow calculations.

The parameterization of the underlying processes in different models may vary, but almost all of them need some degree of parameter calibration against observations to achieve reliable predictions (Beven, 1989; Gong et al., 2015; Troy et al., 2008; Wagener & Gupta, 2005; Yadav et al., 2007). Most of the time, the models are calibrated against observed river streamflow, given its better availability than other training targets like soil moisture or evapotranspiration (Demaria et al., 2007; Gao et al., 2018; Troy et al., 2008; Yadav et al., 2007). Since the beginning, calibration has been an essential effort along with modeling and a continuously evolving process. This is due to the fact that models continue to grow, not only in terms of better parameterization, but also improved spatial resolution, better soil/vegetation data, better/longer meteorological inputs, as well as better training data. On one hand, such new changes can make existing model parameters less optimal. On the other, new developments in river
hydrodynamic modeling have been pushing the size of runoff input units to increasingly finer scales, for instance, the Hillslope River Routing (HRR) model (Beighley et al., 2009) uses 74,075 unit catchments over the conterminous United States (CONUS) and the National Water Model (NWM; http://water.noaa.gov/about/nwm) uses the National Hydrography Dataset Plus Version 2 (NHDPlusV2; http://www.horizon-systems.com/nhdplus/NHDPlusV2_data.php) with ~2.7 million unit river reaches defined over the same region. Recent 1-D hydrodynamic models like the Routing Application for Parallel computation of Discharge (RAPID; David et al., 2011); WRF-Hydro (Gochis et al., 2018), and CaMa-Flood (Yamazaki et al., 2011; Yamazaki et al., 2013) are all implemented at very high spatial resolution, not to mention the 2-D hydrodynamic models which work at even finer scales (Bates et al., 2010; Schumann et al., 2013; Yin et al., 2016; Yin et al., 2017; Yu & Lane, 2006). So there is an urgent need to improve the calibration density of hydrologic models in order to match the need of river hydrodynamic models.

We will illustrate the issues related to calibration further by focusing on the Variable Infiltration Capacity (VIC) model (Liang et al., 1994; Liang et al., 1996), which has been widely implemented over the past decades in numerous large scale studies (Mitchell et al., 2004; Nijssen et al., 2001; Wu et al., 2018; Xia et al., 2012; Zhang et al., 2014) and one major source of runoff inputs to river hydrodynamic modeling (e.g., Lohmann routing of Lohmann et al., 1998; Wen routing of Guo et al., 2004; MOSART of Li et al., 2013; DRTR of Wu et al., 2014; HRR of Ray et al., 2016; CaMa-Flood of Willner et al., 2018). The initial effort to calibrate the VIC model at continental scale started in Maurer et al. (2002), where the VIC soil parameters describing soil depth, baseflow drainage and infiltration capacity were adjusted to best match the monthly streamflow records at the outlets of 12 major river basins/sub-basins in the CONUS. The resulting VIC parameters were used in the North American Land Data Assimilation System phase 1 (NLDAS-1) project (Mitchell et al., 2004), and they are still being used in the operational part of the North American Land Data Assimilation System phase 2 (NLDAS-2) project and many other current practices of continental-scale hydrologic assessments (Brekke et al., 2013; Leng et al., 2016; Mahat et al., 2017; Mizukami et al., 2016). This continental calibration is rather sparse (12 basins) and the parameters form large chunks on the map with strong discontinuities between basin boundaries (Figure 1 (a1), (b1) and (c1), Mizukami et al., 2016; Mizukami et al., 2017), which is a result of the limitations on computing power and data availability at the time. Troy et al. (2008) tried to improve the calibration by training the model over more basins. Assuming the travel time is negligible for small basins at monthly scale and streamflow is fully equivalent to runoff, the monthly runoff data over 1131 small and less-regulated basins collected by United States Geological Survey (USGS) were used to estimate (by interpolation) monthly runoff ratio time series every 1°x1° box over the CONUS. The VIC model was calibrated at every 1°x1° box (~800 of them) during a 4-year period and then the calibrated parameters were downscaled to 1/8° resolution. Troy et al. (2008) considerably improved the spatial coherence of parameters by eliminating the discontinuities between large basins, though artifacts (e.g., bull's eyes, see Figure 1 (a2), (b2) and (c2)) exist due to the coarse sampling (1° spacing) that is insufficient to capture the full spatial variability of runoff. More recently, Oubeidillah et al. (2014) made a very large-scale effort to calibrate the model at a higher resolution of 1/24° against USGS WaterWatch monthly runoff dataset for 2,107 hydrologic sub-basins (8 digit hydrologic units, HUC8s, detailed hydrologic maps can be found at https://water.usgs.gov/GIS/huc.html). This effort significantly increased the density of training runoff data - from 12 (Maurer et al., 2002) and ~800 (Troy et al., 2008) to 2,107 over the CONUS, though parameter discontinuity still exists among HUC8 basins (Figure 1 (a3), (b3) and (c3)). Further increases in the spatial density of training runoff data beyond HUC8s seem unlikely until new techniques like Inverse Streamflow Routing (Fisher et al., 2018; Pan & Wood, 2013) become more mature. Therefore, more recent efforts attempted to relate parameters to soil properties through transfer functions such that the parameter maps can be made at high resolution with good spatial patterns (consistent with soil property maps), for example, the Multiscale Parameter Regionalization (MPR) developed in Samaniego et al. (2010). Mizukami et al. (2017) applied an MPR-type method to VIC calibration over the CONUS using USGS observations over 531 basins and it was found that the skill of MPR was very reasonable but relatively lower than basin-by-basin calibration.

Roughly speaking, two types of efforts have been made to increase calibration density (given available river gauges) and estimate spatially distributed hydrologic parameters:
Regionalize the parameters. The parameter regionalization can be done after (post regionalization) or during the calibration (simultaneous regionalization) through spatial proximity (Oudin et al., 2008; Parajka et al., 2005; Troy et al., 2008; Zhang et al., 2014), climatic and/or physiographic similarity (Beck et al., 2016; Kay et al., 2006; McIntyre et al., 2005; Oudin et al., 2008), and regression between parameters and catchment characteristics (Donnelly et al., 2016; Hundecha & Bárdossy, 2004; Oudin et al., 2008). MPR is a type of simultaneous regionalization but it is also different from others—it regionalizes at the sub-grid scale and then upscales parameters to facilitate the transferability of parameters across scales and locations (Samaniego et al., 2010).

Regionalize the streamflow or its characteristics (e.g., runoff depth, runoff ratio, high and low flow metrics) first then calibrate the model everywhere (Castiglioni et al., 2010; Oubeidillah et al., 2014; Troy et al., 2008; Yadav et al., 2007).

The major difference between the two is which one to regionalize—model parameters or training data, and the skill of calibration will be highly dependent on the efficacy of such regionalization. In some cases, the behavior of the model parameters is better understood (e.g. their relationship with soil properties), and in others the parameter behavior is too elusive for reliable parameter transfer. Recent studies in streamflow climatology analysis (Beck et al., 2015) suggest that the streamflow characteristics (e.g. long-term mean, high/low flows, baseflow decay) can be fairly well interpolated (regionalized) with the help of ancillary data (climate/soil/terrain/vegetation characteristics) even though the actual streamflow time series remain hard to obtain.

Figure 1. Map of VIC model parameters: (a) $b_i$, (b) thick, and (c) $Ds$. Different data sources are labeled with numbers: (1) NLDAS-2; (2) Troy et al. (2008); (3) Oubeidillah et al. (2014); (4) Mizukami et al. (2017); and (5) this study. Background is 2-digit hydrologic unit (HUC2) boundary. The 2-digit numbers represent the HUC2 ID.
to interpolate. Beck et al. (2015) created 1/8° maps of streamflow characteristics globally (the Global Streamflow Characteristics Dataset - GSCD; www.gloh2o.org/GSCD), and the validation suggests a reasonable reliability.

Motivated by the pressing need to significantly improve the calibration density (and calibration period length) and inspired by Beck et al. (2015) work, this study will calibrate VIC model at 1/8° resolution against streamflow characteristics over the CONUS, increasing the calibration density from 2,107 HUC8 basins (Oubeidillah et al., 2014) to 52,663 grid-cells over the CONUS, for a 20-year training period and 10-year validation period. The paper is structured as follows: Section 2 describes the data source and methodology used to calibrate and validate the parameters. The calibration performance and the improvements to calibration strategy are presented in Section 3. Finally, conclusions are provided in Section 4.

2. Methodology
2.1. Parameter calibration strategy

We set up the VIC model across the CONUS at the same 1/8° resolution as the GSCD data and calibrated each 1/8° grid-cell (a total of 52,663 grid-cells) independently on streamflow characteristics. The GSCD provides global maps of 17 streamflow statistics including the mean flow, different high/low flow percentiles (flow duration curve - FDC), various baseflow indices, and timing of half flow volume of the year. Such characteristics, though less informative than the full time series, can help constrain many important aspects of the hydrograph. To derive the GSCD, daily streamflow observations from 3,000 to 4,000 catchments (10–100,000 km²) worldwide were used to train artificial neural network ensembles to estimate streamflow characteristics based on 20 climatic and physiographic characteristics for the catchments. The trained neural network ensembles were subsequently applied spatially over the entire ice-free land surface, resulting in global maps of the streamflow characteristics. More details can be found in Beck et al. (2013, 2015). Note that the GSCD uses the term “streamflow” for the 1/8° gridded dataset, while the modeling literature normally use the term “runoff” for the same gridded quantity and “streamflow” for the flow values routed to the outlets of river basins (i.e. runoff integrated in time and space). The two terms are equivalent at the grid-scale in this study since the routing processes within the grid-cells can be ignored at the daily scale, and thus they will be used interchangeably here for grid-scale quantities. The term “streamflow” will be used for flows routed to outlets of larger basins.

There are a variety of typical streamflow characteristics proven to be useful for model calibration, including the FDC, baseflow indices, runoff ratio, master recession curve (Addor et al., 2018; Fenicia et al., 2018; Troy et al., 2008; Winsemius et al., 2009; Yilmaz et al., 2008). Moreover, the experiments conducted by Fenicia et al. (2018) showed that calibration against several FDC quantiles and baseflow index can produce virtually indistinguishable streamflow prediction from calibration operated directly on the streamflow time series. Therefore, the four most significant streamflow characteristics were selected as the calibration targets, namely, mean annual streamflow per unit area (QMEAN; mm/yr), baseflow index (BFI in this study, BF13 in the GSCD; dimensionless), high flow (Q10, the daily flow exceeded 10% of the time; mm/d), and low flow (Q90, the daily flow exceeded 90% of the time; mm/d), to constrain different aspects of the hydrograph.

Given the selected streamflow characteristics and VIC model parameter sensitivity analysis by Demaria et al. (2007), three VIC parameters controlling the generation of surface and subsurface flow were selected for calibration (Table 1), including the variable infiltration curve parameter \( b_i \), thickness of soil layer 2 \( \text{thick}_2 \), and fraction of the maximum velocity of base flow where nonlinear base flow begins \( D_s \). This is a limited set of parameters to tune, as constrained by the number of calibration targets (four streamflow characteristics), but they are still powerful enough to help adjust most model behaviors.

The Shuffled Complex Evolution (SCE-UA) algorithm (Duan et al., 1992, 1994) was employed to find the optimal parameter set for each grid-cell. The objective function for each grid-cell was:

| Parameter | Range  | Units | Description |
|-----------|--------|-------|-------------|
| \( b_i \) | 0.0001-0.5 | N/A | Variable infiltration curve parameter |
| \( \text{thick}_2 \) | 0.1-2 | m | Thickness of soil layer 2 |
| \( D_s \) | 0–1 | N/A | Fraction of the maximum velocity of base flow where nonlinear base flow begins |
\[ \text{Obj} = w_0 \ln \left( \frac{Q_{\text{MEAN}_m}}{Q_{\text{MEAN}_o}} \right) + w_1 \ln \left( \frac{\text{BFI}_m}{\text{BFI}_o} \right) + w_2 \ln \left( \frac{Q_{10m}}{Q_{10o}} \right) + w_3 \ln \left( \frac{Q_{90m}}{Q_{90o}} \right) \]  

(1)

where \(Q_{\text{MEAN}_m}, \text{BFI}_m, Q_{10m}, Q_{90m}\) were GSCD values, and \(Q_{\text{MEAN}_o}, \text{BFI}_o, Q_{10o}, Q_{90o}\) are calculated from the model simulation. \(w_i(i = 1, 2, 3, 4)\) represents the weighting factor, indicating the importance of the corresponding characteristic during calibration. Since the uncertainty behaviors and their definitions in GSCD are different for the four characteristics, they cannot be directly used to determine the weights. Moreover, the weight allocation itself is inevitably a more subjective process, which should be customized according to the specific focus or needs of the study. With considerations on the focus of this work on improving model overall performance and a simple assessment of their relative importance (subjective), \(w_i(i = 1, 2, 3, 4)\) was set to \([0.6, 0.2, 0.1, 0.1]\) for the four characteristics in this study.

### 2.2. Model set-up, calibration, and evaluation

The VIC model (version 4.0.6, https://github.com/yuanyangthu/VIC4.06_netcdf) simulation was performed at daily time steps in water balance mode (Liang et al., 1994), forced by the meteorological data from Livneh et al. (2013), including daily minimum and maximum temperature, precipitation and wind speed. The initial values for all model soil and vegetation parameters (whether to calibrate or not) were obtained from gridded data sets developed for the NLDAS-2 (Xia, Mitchell, Ek, Sheffield, et al., 2012).

The model ran at two different resolutions during the calibration and evaluation. The calibration was performed at 1/8° for the 20-year period of 1982-2001, with 2 years of model spin-up during 1980-1981. The model evaluation was performed at a slightly higher resolution of 1/16° for the 10-year period of 2002-2011. The model was also evaluated for the calibration period 1982-2001. This higher resolution was used for two reasons: (1) the meteorological forcing data (Livneh et al., 2013) has its original resolution at 1/16°, (2) finer spatial resolution of the runoff can better match the unit catchments used by the routing model. After the parameters were calibrated at 1/8°, they were resampled to 1/16° using bilinear interpolation before evaluation.

To evaluate the VIC performance against observed streamflow at the outlets of larger basins, the modeled grid-cell runoff (surface and subsurface) was re-mapped and routed using a topographically-based routing model, the HRR model (Beighley et al., 2009). The model is built on an irregular model unit defined from the river network and associated catchment boundaries. For each model unit (catchment over a river reach), the landscape is approximated as an open book with two planes draining laterally to a main channel. The kinematic wave method is used for lateral transport of surface and subsurface runoff, while the diffusion wave (Muskingham-Cunge) method is used for river routing. The model has been applied in many regions, such as the Amazon (Beighley et al., 2009), Congo (Beighley et al., 2011), Susquehanna (Ray et al., 2016; Seyyedi et al., 2014, 2015) and Ohio River Basin (Zhao & Beighley, 2016). In this application, the NLDAS domain was delineated into 74,075 river reaches and catchments based on a derived river network using a threshold area of roughly 100 km². The resulting drainage area for the model units ranges from 0.1 to 2,700 km², with a median value of 120 km². The hillslope, channel and floodplain length and gradients required by the HRR model were determined using the 3 arc-sec (~90 m) elevation and flow direction dataset derived from NASA’s Shuttle Radar Topography Mission (SRTM; Lehner et al., 2008). The main routing parameters, hillslope surface roughness, subsurface horizontal conductivity, and channel roughness were derived following Zhao and Beighley (2016).

The observed daily streamflow data in this study were part of the Geospatial Attributes of Gages for Evaluation Streamflow, version II (GAGES II; Falcone et al., 2010), and were downloaded from USGS website (https://waterdata.usgs.gov/nwis/dv?referred_module=sw). Two groups of catchments were used for the evaluation. The first group of catchments (Group 1) were those used in deriving the characteristics in the GSCD, and the second group of catchments (Group 2) were those not used in deriving GSCD, which were used to test the robustness of the calibration method, especially in ungauged regions. In addition, these catchments were required to have more than 365 days of valid data (not necessarily consecutive) during calibration/validation period, resulting in 704 (calibration period)/648 (validation period) catchments in Group 1 and 3941 (calibration period)/3809 (validation period) in Group 2 that were suitable for the evaluation. These catchments are small-to-medium-sized catchments ranging from 70 to 26,000 km². Additionally,
we also selected five large and not heavily regulated basins which are larger than 26,000 km² and have a 30-year complete data record from 1982-2011 to complement the model evaluation. Four skill metrics, the relative bias (RB), alpha, correlation coefficient (CC) and Nash-Sutcliffe Efficiency (NSE; Nash & Sutcliffe, 1970) coefficient, were used to evaluate model performance. RB provides a measure on the total streamflow volume error. Alpha is a measure of relative variability in the simulated and observed values. CC tests the agreement between the simulated streamflow dynamics and gauge observations. NSE measures the overall predictive skill of the model as compared to the mean daily observed streamflow. Besides, another four similar metrics, including histogram intersection (histo match), spatial correlation coefficient (spatial CC), spatial variability error (CV ratio) and SPAtial Efficiency (SPAEF), were introduced to evaluate the spatial pattern of parameters and streamflow characteristics (Demirel et al., 2018; Koch et al., 2018). The metrics are summarized in Table 2.

### 3. Results and Discussions

#### 3.1. Spatial patterns of calibrated parameters

Figure 1 shows the spatial distribution of the three calibrated VIC parameters from five data sources: (1) NLDAS-2; (2) Troy et al. (2008); (3) Oubeidillah et al. (2014); (4) Mizukami et al. (2017); and (5) this study. Note that Mizukami et al. (2017) calibrated the parameters in Nijssen baseflow formulation instead of the Arno baseflow formulation, therefore Ds shown in Figure 1 (c4) was the Arno Ds converted from Nijssen baseflow parameters. The calibration methods to derive these five sets of parameters can be divided into two categories: (1) non-MPR approach, which directly calibrates parameters against observations; and (2) MPR approach, which relates parameters to basin predictors (e.g., elevation, slope, soil texture, or vegetation characteristics) through transfer function and then calibrates these “transfer function parameters”. The parameters from Mizukami et al. (2017), derived by the MPR approach, were functions of topography and soil properties and showed very good spatial patterns (Figure 1 a4, b4, and c4). Among the other four sets of parameters, which were derived by non-MPR approach, the parameters calibrated in this study exhibited a much better spatial coherence, without discontinuities across basin boundaries or bull’s eye patterns from simple parameter interpolation. For example, the topography, climate and soil characteristics are similar in the Rocky Mountains, which indicates that parameters would not vary significantly. However, the parameters from NLDAS-2 located in the same region or subregion of...
No. 10, 14, 16 and 17 HUC2 basins changed abruptly from one region/subregion to another. Such patchiness also existed in calibrated parameters from Oubeidillah et al. (2014), but across HUC8 basins. Parameters in Troy et al. (2008) exhibited some clusters with high values in the center and low values around. Compared with the former three sources of parameters, the calibrated parameters in this study varied much more consistently in space. The parameter $b_i$ controls the shape of the variable infiltration curve in VIC, and effectively dictates the partitioning of rainfall into infiltration and surface runoff. Smaller $b_i$ results in an increase in the infiltration capacity and thus a decrease in the amount of surface runoff. Humid regions typically had a higher $b_i$ while arid regions had a lower $b_i$, as shown in Figure 1 (a5). In contrast, the parameter thick$_2$ showed an opposite pattern with higher values in arid regions and lower values in humid regions (Figure 1 (b5)) since thicker soil depths can retain more water and increase the loss due to evapotranspiration. The spatial patterns of the calibrated $b_i$ and thick$_2$ match well with the spatial patterns in aridity and runoff ratio from the CAMELS (Catchment Attributes and MEteorology for Large-sample Studies) data set (Addor et al., 2017), as shown in Figure 2.

$b_i$ was significantly and negatively correlated with aridity, with a spatial correlation of $-0.587$ ($P<0.001$), while positively correlated with runoff ratio, with a correlation of 0.416 ($P<0.001$). In contrast, thick$_2$ was significantly and positively correlated with aridity (spatial CC = 0.434, $P<0.001$), negatively correlated with runoff ratio (spatial CC = -0.418, $P<0.001$). $Ds$ is a more conceptual parameter controlling the baseflow rate. It showed no obvious relationship with climate, topographic, soil, or vegetation characteristics and was more likely to be the result of more complex interactions between model parameters and streamflow characteristics.

3.2. Assessment of the grid-level distributed calibration strategy

The performance of the calibrated model was compared to that of the pre-calibrated model (i.e., the same model setups except using NLDAS-2 parameters in Figure 1 (a1), (b1) and (c1)) to assess the effectiveness of this calibration. NLDAS-2 parameters were used as benchmark for skill evaluation since they have been used for many studies. Figure 3 shows cumulative density function (CDF) plots for the RB, alpha, CC, and NSE values for the catchments of Group 1 for both the calibration (solid) and validation (dashed) periods. All metrics were computed based on daily VIC model simulations before calibration (blue lines) and after calibration (red lines). Tables in Figure 3 show the statistics of Kolmogorov-Smirnov test (K-S test; Eghbali, 1979; Smirnov, 1948), which was used to validate whether the improvements were significant. Significant improvements can be seen in RB as evidenced by steeper CDFs as well as median values that were closer to zero (Figure 3a). For the calibration period, the percentages of RB falling in ranges < -25%, [-25%, 25%] and > 25% were 7.2%, 75.5% and 17.3%, improved from 16.6%, 56.8% and 26.6%, respectively. This indicated that the large underestimation and overestimation was reduced after calibration. The calibration tended to underestimate the flow variability to a greater degree, which, at the same time, reduced the overestimation of the flow variability in pre-calibration (Figure 3b). Therefore, higher percentages of alpha fell in ranges around 1.0, such as the range [0.5, 1.5]. As expected, there was no obvious improvement in terms of CC since the streamflow characteristics used to calibrate model did not provide much information about the temporal dynamics (Figure 3c). In terms of NSE, the model performances were improved as indicated by the

Figure 2. Map of (a) aridity and (b) runoff ratio from CAMELS data set.
rightward shift of the red lines in Figure 3d. The NSE changed from negative to positive for approximately 10% of catchments for the calibration period. About 50% of the catchments after calibration had an NSE greater than 0.4.

To study the spatial pattern of model performance, the skill metrics were plotted at each catchment of Group 1 for both the calibration and validation periods in Figure 4. The improvements in model performance varied among different parts of the CONUS. The relative bias has been reduced, especially for the eastern part of the CONUS. The high positive streamflow biases in the eastern Great Plains and the southeastern CONUS have been reduced to less than ±30% for the majority of the catchments. The negative biases around the Great Lakes region have been reduced to less than ±10%. Improvements can also be seen elsewhere, for example in the northeast. Basins west of 100°W showed roughly the same strong positive biases after calibration, and thus exhibited little improvement. The change of the flow variability error showed a different spatial pattern, with overestimation being largely reduced in the western CONUS and underestimation being worsened in some regions, such as Texas, Kansas and Missouri. In terms of CC, there were no obvious changes after calibration. The change of NSE did not show an obvious spatial pattern.

Similar results can be seen for catchments which are not used to derive the GSCD (Figure S1 and S2). For example, the improvements in RB were still concentrated in the eastern CONUS (Figure S2). The strong overestimation in the southeastern CONUS and the eastern Great Plains has been reduced. The underestimation in the northeast has also been reduced.
Figure 4. Map of relative bias (RB), alpha, correlation coefficient (CC), and Nash-Sutcliffe Efficiency (NSE) coefficient over catchments of Group 1 for calibration (left column) and validation (right column) periods. Metric values before and after calibration are labeled by numbers 1 and 2.
Clearly, while the calibrated parameter set may provide satisfactory results for wetter regions, it is challenging to capture the streamflow in drier regions (Figure 4), which is consistent with previous studies (Mizukami et al., 2017; Naz et al., 2016; Oubeidillah et al., 2014; Xia, Mitchell, Ek, Sheffield, et al., 2012). Figure 5 shows model metrics as a function of runoff ratio (RC, obtained from GSCD) for each small-to-medium sized catchment. It shows a tendency for VIC to perform worse for catchment with a lower runoff ratio than catchments with higher runoff ratios. Possible reasons for poor performance in the dry region included errors in forcing data (Beck et al., 2016; Lohmann et al., 2004; Mizukami et al., 2017; Naz et al., 2018; Oubeidillah et al., 2014; Troy et al., 2008; Xia et al., 2012), anthropogenic effects on streamflow (Maurer et al., 2002; Troy et al., 2008; Xia, Mitchell, Ek, Cosgrove, et al., 2012) as well as limitations in model physics (Demaria et al., 2007; Naz et al., 2018; Newman et al., 2017; Oubeidillah et al., 2014). Issues in dry regions are being addressed in newer versions of VIC model. For example, the “Big Leaf” model, which leads to evapotranspiration (ET) underestimation due to lack of soil evaporation between plants, has been replaced by the “Clumped” model that considers such soil evaporation (see VIC model documentation at https://vic.readthedocs.io/en/master/Overview/ModelOverview/).

Three typical catchments from the southeastern CONUS, the southwestern CONUS and the northern Rocky Mountains were selected to showcase the region-specific model performance and efficacy of the calibration method. Figure 6 shows daily hydrographs for the three catchments for one year (2001) in the calibration period and one year (2002) in the validation period. The catchment Little River near Silverstreet, South Carolina (Figure 7), representative of the southeastern CONUS, exhibited much better model performance after calibration (Figure 6a). The simulation before calibration could capture the daily variability, but produced too much streamflow, especially for the dry season. The calibration significantly reduced the overestimation and showed a good match with the observations for this humid subtropical catchment. Arid regions have always been a challenge for the VIC model. For many catchments in these regions, the calibration was...
only able to reduce the large wet bias in the model to a certain extent and the post-calibration bias was still too high. For example, Gila River near Gila, New Mexico, which locates in the southwestern CONUS and is a typical arid catchment, exhibited much higher streamflow simulations than the observations both before and after calibration (Figure 6b). The calibration in this study, which did not calibrate snow parameters, did not work well in snowmelt-dominated region. Figure 6c shows the observed and modeled streamflow at Bull Lake Creek above Bull Lake, Wyoming. The station is 1790 m above sea level and therefore the runoff is highly influenced by snowmelt. The modeled snowmelt seemed to have started too early, causing peak flows during April, while the observed peaks arrived in May. This early snowmelt is mainly affected by the one-layer snow band set in VIC and indicates that multiple snow bands should be calibrated and implemented. The VIC model both before and after calibration also underestimated streamflow at this station. This negative bias can be attributed to forcing errors (Pan et al., 2003; Sheffield et al., 2003) and inefficiency in snowpack related parameterizations and parameters (Xia et al., 2018).

The GSCD was derived entirely using data from small basins. To complement the previous evaluation using the small basins, we performed a fully independent evaluation of the calibration strategy over five large basins (Figure 7). Table 3 shows the model performance for the five large basins at daily, monthly and yearly scales. To give an emphasis on relative bias, which measures the model skills in simulating total amount of

Figure 6. Daily streamflow at the USGS stations (a) Litter River near Silverstreet, South Carolina, USGS 02167450; (b) Gila River near Gila, New Mexico, USGS 09430500; and (c) Bull Lake Creek above Bull Lake, Wyoming, USGS 06224000.

Figure 7. The location of the three selected USGS stations and five large basins.
runoff, annual streamflow is shown in Figure 8. As shown in Table 3 and Figure 8, the total amount of runoff was well simulated after calibration. The Potomac River, which showed a good performance prior to the calibration, continued to perform well. The Alabama River, for which the model overestimated streamflow, showed marked improvements during both the calibration and validation periods, with relative bias reduced to -4.11% and 8.51% from 24.74% and 40.32%, respectively. The calibrated model accurately reproduced the magnitude and variability of annual streamflow. Notable improvements were also observed in the Ohio River and Mississippi River, where the underestimation was strongly reduced. The Arkansas River, which is located in the arid region, showed slightly larger relative biases after calibration. Since the pre-calibration

| Basins   | Calibration period | Validation period |
|----------|--------------------|-------------------|
| Potomac  | -1.26              | 4.62              |
| Alabama  | 24.74              | 40.32             |
| Ohio     | -10.63             | -4.76             |
| Mississippi | -23.86          | -16.77            |
| Arkansas | 13.66              | 18.71             |

Table 3

Model Performance for the Eight Large Basins at Daily, Monthly and Yearly Scales

| Basins   | Calibration period | Validation period |
|----------|--------------------|-------------------|
| Potomac  | -1.26              | 4.62              |
| Alabama  | 24.74              | 40.32             |
| Ohio     | -10.63             | -4.76             |
| Mississippi | -23.86          | -16.77            |
| Arkansas | 13.66              | 18.71             |

Figure 8. Annual streamflow time series at five large basins in Figure 7 for the period 1982-2011.
parameters values (NLDAS-2) were already calibrated to maximize NSE in large basins (Maurer et al., 2002), the calibration here did not show much improvements in NSE, especially at daily and monthly scales. The improvements were mostly witnessed by basins which did not perform well prior to the calibration, such as the Alabama River and Mississippi River. Overall, the model performance in large basins after calibration was good, which is inspiring as it is an independent validation of our grid calibration strategy and shows that it is comparable to the large basin calibration approach.

The model performance in the validation period was comparable to that in the calibration period in most catchments (Figure 3, 4, 8 and Table 3). Overall, this indicated that the model parameters were robust enough across different periods with comparable skills.

3.3. Spatial distribution of QMEAN, Q10, Q90, BFI

Figure 9 shows spatial maps of the four streamflow characteristics used for the calibration. For each characteristic, we compared four sources, including the GSCD and maps derived from the three simulations (2002-2011) based on NLDAS-2 parameters, Oubeidillah et al. (2014) parameters and parameters calibrated in this study. Note that we directly used the daily runoff simulations from Oubeidillah et al. (2014), considering that the model performed best and their parameters were most valid in the original setup (model version, forcing data and spatial resolution). The spatial pattern of QMEAN after the calibration performed in this study visually corresponded best with the GSCD. The overestimation in the southeastern CONUS and Great Plains and the underestimation in the northeastern CONUS and Great Lakes region by NLDAS-2 have been reduced. In terms of BFI, the GSCD and the map derived from the calibration in this study showed good agreement and a much better spatial coherence than the maps derived from NLDAS-2 and Oubeidillah et al. (2014) calibrations, which showed pronounced discontinuities across large basins and HUC8 basins, respectively. The BFI overestimation by NLDAS-2 in the east has been reduced significantly. The spatial patterns in BFI from the GSCD and the map derived here were quite similar to those from the CAMELS dataset (see Figure 4e of Addor et al., 2017), although different methods were used for the BFI calculation. The baseflow in CAMELS dataset was derived by the Ladson et al. (2013) digital filter, while the baseflow in the GSCD was the 7-day moving minimum streamflow. Compared with the GSCD, the BFI map after calibration had higher BFI in the most regions of the southwestern CONUS, which means the parameters calibrated in this study produced too much slow runoff. The spatial distribution of Q10 and Q90 followed closely the pattern of QMEAN. As was the case of QMEAN, calibration raised both the high flow and low flow in the northeast and Great Lakes region, lowered them in the southeast and Great Plains. However, there were still some large differences between the maps of GSCD and calibrated simulation. For example, the calibrated simulation produced higher Q10 in the area covering Nevada, Arizona, New Mexico and Texas, and higher Q90 in Missouri and Illinois. The limited improvement in Q10 and Q90 may due to the lower weights assigned to them in the objective function.

Table 4 further shows the quantitative spatial comparison of the four streamflow characteristics before and after calibration. All the four streamflow characteristics derived here showed higher spatial correlations with those from GSCD. The CV ratio of QMEAN, BFI, and Q90 increased after calibration, showing the calibration improved the spatial variability of these streamflow characteristics. Except the slight decrease in QMEAN (0.988 to 0.980), the streamflow characteristics after calibration witnessed increases in histo match, which means that flow value distributions in different ranges were better reproduced. In terms of SPAEF, a comprehensive metric considering the above three metrics together and providing a more holistic and balanced assessment for spatial performance, the SPAEF of QMEAN, BFI, Q10 and Q90 increased to 0.916, 0.571, 0.807 and 0.785 from 0.878, 0.233, 0.780, and 0.489, respectively, showing that the calibration contributed to better spatial patterns for the four streamflow characteristics.

Overall, the streamflow characteristics after calibration were much closer to the GSCD, demonstrating the effectiveness of the calibration strategy. But it should be noted that perfect agreement between the characteristic maps of GSCD and the VIC-based maps was unlikely, since (1) calibration had some limitations in some regions and could not fix the problem caused by model physics or forcing errors and (2) the observed streamflow data used by GSCD (>10 years but not necessarily consecutive after selection) covered a different time period from the VIC-based simulated streamflow data (10-year period covering 2002-2011).
Figure 9. The four streamflow characteristics: (a) QMEAN; (b) BFI; (c) Q10 and (d) Q90. Different data sources are indicated by numbers: (1) GSCD, simulations based on (2) NLDAS-2 parameters, (3) Oubeidillah et al. (2014) parameters, and (4) parameters calibrated in this study.
3.4. Advantages and disadvantages of the calibration strategy

The grid-level distributed calibration against GSCD streamflow characteristics provided a highest possible calibration density in space for large-scale hydrologic modeling. This calibration strategy can also be applied in sparsely gauged or ungauged regions. It has been applied in Lin et al. (2019) for global discharge reconstruction. It is applicable to any hydrologic model and does not try to establish relationships between conceptual model parameters and watershed characteristics, which need to be reestablished when calibrating different models. Moreover, directly constraining the model to reproduce several hydrologically important characteristics of the hydrograph can make the calibration more robust to deficiencies in simple objective functions of streamflow time series since these characteristics are less sensitive to features poorly reproduced by the hydrologic model, such as the errors in the timing and magnitude of several individual flood peaks (Fenicia et al., 2018).

Although this calibration method offers appealing benefits, limitations also exist. First, the use of streamflow characteristics is usually associated with a loss of information in lieu of the streamflow time series, which may have a negative impact on parameter estimation (Fenicia et al., 2018; Shaﬁi & Tolson, 2015). The “adequacy” of available characteristics requires further investigation. This study relied on only four streamflow characteristic values provided by GSCD. Not much information in time dimension except the high flow (Q10) and low flow (Q90) was involved, which may be responsible for the lack of improvements in daily dynamics (Figure 3c and Figure 4). While monthly streamflow characteristics could help to solve this issue, more uncertainties would be encountered when developing the monthly streamflow characteristics. More streamflow characteristics, such as the flashiness index used in Fenicia et al. (2018), could be considered in the calibration process to improve the variability and dynamics of the streamflow modeling. Additionally, the quality of streamflow characteristics data matters too. GSCD was shown to be very reliable over CONUS thanks to the dense streamflow monitoring networks in the region (Beck et al., 2015), while extra caution should be exercised in global applications.

Second, this calibration strategy and resulted parameters are inevitably forcing-dependent, because the calibrated parameters can compensate forcing biases and model structure problems (Blöschl et al., 2007; Bock et al., 2016; Elsner et al., 2014; Merz et al., 2011; Weiland et al., 2015). This would reduce the skill of the model in replicating results outside of calibration conditions and increase the potential for equifinality of parameter sets and higher model uncertainty. For example, Elsner et al. (2014) demonstrated that the choice of forcing dataset resulted in different parameter combinations, although with similar objective function values. While such forcing dependency is acceptable in the CONUS, given the high quality of the forcing data, this can be an issue in regions with low quality forcing data.

Third, the potential performance improvement will depend on the number of calibrated parameters (Beck et al., 2016). The most sensitive parameters for runoff generation in VIC are mostly “empirical” parameters, i.e., curve shapes, layer thickness and baseflow linearity threshold. Newman et al. (2017) has pointed out that increasing the number of calibrated parameters could improve VIC model performance. Three soil-related parameters were calibrated here. The other soil-related conceptual parameters, such as the maximum base flow velocity and thickness of soil layer 3, could also affect the hydrologic responses and be calibrated rather than estimated a priori using soil texture data or expert opinion. Parameters related to vegetation characteristics could be calibrated as well. Vegetation parameters allow substantially greater

| Table 4 | Spatial Performance for the Four Streamflow Characteristics before and after calibration |
|---------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|
|         | CC                                | CV ratio                          | Histo match                       | SPAEF                            |
| QMEAN   | Before (NLDAS-2)                  | 0.929                            | 0.901                            | 0.988                            | 0.878                            |
|         | After                             | 0.952                            | 0.935                            | 0.980                            | 0.916                            |
| BFI     | Before (NLDAS-2)                  | 0.258                            | 0.925                            | 0.823                            | 0.233                            |
|         | After                             | 0.585                            | 0.949                            | 0.903                            | 0.571                            |
| Q10     | Before (NLDAS-2)                  | 0.884                            | 0.865                            | 0.870                            | 0.780                            |
|         | After                             | 0.929                            | 0.837                            | 0.924                            | 0.807                            |
| Q90     | Before (NLDAS-2)                  | 0.720                            | 0.595                            | 0.862                            | 0.489                            |
|         | After                             | 0.850                            | 0.857                            | 0.943                            | 0.785                            |
control over the water balance since they can affect directly on ET rates. For example, Troy et al. (2008) calibrated the stomatal resistance factor and Newman et al. (2017) calibrated minimum stomatal resistance and monthly varying LAI to achieve better model performance. However, at the same time, we need to bear in mind that increasing the number of free calibration parameters may also increase the equifinality problem. Extra attention is needed while calibrating more parameters.

Fourth, the resolution of calibration targets (streamflow characteristics) limits calibration density and thus parameter resolution, a disadvantage against MPR. It would be more difficult to calibrate with this method at a higher resolution because of higher computation cost and potentially higher uncertainty in estimates of gridded streamflow characteristics. Here we calibrated VIC model at 1/8° and then applied the calibrated parameter set at 1/16° after bilinear interpolation. Figure S3 shows model performance for catchments of Group 1 at the two resolutions. Significant differences were found in the distribution of relative bias (Figure S3 (a)). In general, downscaling parameters from 1/8° to 1/16° produced more runoff. This is because at coarser resolutions, the precipitation intensity is lower due to averaging (Troy et al., 2008) – a typical scaling issue. This can be amended if we can downscale runoff properly. The relative biases between 1/8° and 1/16° differed a lot in the western mountainous region (Figure 10), where model with the downscaled 1/16° parameters showed much larger positive biases. The actual runoff differences between 1/8° and 1/16° were not so large in this region, but the small runoff magnitude inflated the relative bias. As for alpha and correlation, their distributions at the two resolutions were very close. Overall, there were no significant differences in the distribution of NSE, indicating that scaling parameters from 1/8° to 1/16° had little effect on model performance in this study. More efforts are needed in the future to improve the scale consistency of distributed parameter estimations.

4. Conclusions

As large-scale hydrologic modeling at fine resolutions is becoming more common and necessary in hydrologic applications, spatially distributed parameters need to be improved. This study introduced a grid-level distributed calibration method using machine learning-based streamflow characteristics for the VIC model over CONUS. The main findings are as follows:

1. A set of distributed parameters at 1/8° resolution was produced with good spatial coherence. The spatial patterns in calibrated parameter values corresponded well with spatial patterns in climate characteristics, such as aridity and runoff ratio.
2. The quality and applicability of the calibrated parameters were validated by better model performance with respect to the observed streamflow from two groups of small-to-medium-sized catchments.
(70-26,000 km², 704/648 (calibration/validation period) in the group used to derive the GSCD, 3941/3809 (calibration/validation period) in the group not used to derive the GSCD) as well as five large basins. Considerable improvements can be seen in relative bias, with large overestimation and underestimation in the east of the CONUS being reduced. However, the catchments in the central and southwest US tended to improve less, suggesting additional issues to solve that are beyond the reach of parameter calibration, e.g., forcing error, anthropogenic effects and model physics.

3. The comparisons on streamflow characteristics during validation period showed that the calibration was able to match these characteristics from the model to GSCD, which further confirmed the efficacy of the calibration scheme.

Overall, this study offered a proof-of-concept for the grid-level distributed parameter calibration at continental scale. As the new calibration strategy enabled a much higher calibration density, the benefits are obvious for high-resolution hydrologic and river hydrodynamic modeling. Besides the intrinsic limitations of calibration measures, the lack of temporal dynamics information also limits the power of this strategy. Future work will extend the experiments to global scale.

Acknowledgments
We thank Naoki Mizukami from National Center for Atmospheric Research for providing calibrated VIC parameters using MPR-flex. We are grateful to Martyn P. Clark, Luis Samaniego and three anonymous reviewers for their insightful comments and constructive criticisms, which helped us substantially improve the study and refine the paper. This work was financially supported by the National Natural Science Foundation of China (Grant NO. 714610107). Hylke E. Beck was supported by the U.S. Army Corps of Engineers' International Center for Integrated Water Resources Management (ICIWaRM), under the auspices of UNESCO. Support was also provided through NASA Grant NNX15AD48G (Deriving Global High Resolution Surface Runoff Fields from Space-borne Altimetry Measurements and Development of Precursor Product from River Gauges). Additional support came from a scholarship from the China Scholarship Council (CSC). The calibrated VIC model parameters over CONUS are permanently archived at DataSpace at Princeton University under the section for Civil and Environmental Engineering, online accessible at http://arks.princeton.edu/ark:88435/dsp010049g519d (search for "Calibrated VIC Model Parameters over CONUS").

References
Addor, N., Nearing, G., Prieto, C., Newman, A. J., le Vine, N., & Clark, M. P. (2018). A ranking of hydrological signatures based on their predictability in space. Water Resources Research, 54(11), 8792–8812. https://doi.org/10.1029/2018WR022606
Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017). The CAMELS data set: catchment attributes and meteorology for large-sample studies. Hydrology and Earth System Sciences, 21(10), 5293–5313. https://doi.org/10.5194/hess-21-5293-2017
Asadieh, B., & Krakauer, N. Y. (2017). Global change in streamflow extremes under climate change over the 21st century. Hydrology and Earth System Sciences, 21(11), 5863–5874. https://doi.org/10.5194/hess-21-5863-2017
Bates, P. D., Horritt, M. S., & Hewitson, T. J. (2010). A simple inertial formulation of the shallow water equations for efficient two-dimensional flood inundation modelling. Journal of Hydrology, 387(1-2), 33-45. https://doi.org/10.1016/j.jhydrol.2010.03.027
Beck, H. E., de Roo, A., & van Dijk, A. I. (2015). Global maps of streamflow characteristics based on observations from several thousand catchments. Journal of Hydro Meteorology, 16(4), 1478–1501. https://doi.org/10.1175/JHM-D-14-01551
Beck, H. E., van Dijk, A. I., de Roo, A., Miralles, D. G., McVicar, T. R., Schellekens, J., & Bruijnzeel, L. A. (2016). Global-scale regionalization of hydrologic model parameters. Water Resources Research, 52, 3599–3622. https://doi.org/10.1002/2015WR018247
Beck, H. E., van Dijk, A. I., Miralles, D. G., de Jeu, R. A., Bruijnzeel, L. A., McVicar, T. R., & Schellekens, J. (2013). Global patterns in base flow index and recession based on streamflow observations from 3334 catchments. Water Resources Research, 49, 7843–7863. https://doi.org/10.1002/wrcr.2013918
Beighley, R., Ray, R. L., He, Y., Lee, H., Schaller, L., Andreadis, K. M., et al. (2011). Comparing satellite derived precipitation datasets using the Hillslope River Routing (HRR) model in the Congo River Basin. Hydrological Processes, 25(20), 3216–3229. https://doi.org/10.1002/ hyp.5045
Beighley, R. E., Eggert, K. G., Dunne, T., He, Y., Gummadi, V., & Verdin, K. L. (2009). Simulating hydrologic and hydraulic processes throughout the Amazon River Basin. Hydrological Processes, 23(8), 1221–1235. https://doi.org/10.1002/hyp.7252
Beven, K. (1989). Changing ideas in hydrology—the case of physically-based models. Journal of Hydrology, 105(1-2), 157–172. https://doi.org/10.1016/0022-1694(89)90010-7
Blöschl, G., Ardon Bardin, S., Bonell, M., Dörning, M., Gobrecht, D., Gutknecht, D., et al. (2007). At what scales do climate variability and land cover change impact on flooding and low flows? Hydrological Processes: An International Journal, 21(9), 1241–1247. https://doi.org/10.1002/hyp.6669
Bock, A. R., Hay, L. E., McCabe, G. J., Markstrom, S. L., & Atkinson, R. D. (2016). Parameter regionalization of a monthly water balance model for the conterminous United States. Hydrology and Earth System Sciences, 20(7), 2861–2876. https://doi.org/10.5194/hess-20-2861-2016
Brekke, L., Thrasher, B. L., Maurer, E. P., & Pruitt, T. (2013). Downscaled CMIP3 and CMIP5 climate projections: release of downscaled CMIP5 climate projections, comparison with preceding trends, and summary of user needs. US Department of the Interior, Bureau of Reclamation. Technical Service Center, Denver, Colorado, 116.
Castiglioni, S., Lombardi, L., Toth, E., Castellarin, A., & Montanari, A. (2010). Calibration of rainfall-runoff models in ungauged basins: A regional maximum likelihood approach. Advances in Water Resources, 33(10), 1235–1242. https://doi.org/10.1016/j.advwatres.2010.04.009
David, C. H., Maidment, D. R., Niu, G.-Y., Yang, Z.-L., Habets, F., & Eijkhout, V. (2011). River network routing on the NHDPlus dataset. Journal of Hydrometeorology, 12(5), 913–934. https://doi.org/10.1175/2011JHM1345.1
Demaria, E. M., Nijsen, B., & Wagener, T. (2007). Monte Carlo sensitivity analysis of land surface parameters using the Variable Infiltration Capacity model. Journal of Geophysical Research, 112, D11113. https://doi.org/10.1029/2006JD007534
Demirel, M. C., Mai, J., Mendiguren, G., Koch, J., Samaniego, L., & Sitos, S. (2018). Combining satellite data and subjective objective models for improved spatial pattern performance of a distributed hydrologic model. Hydrology and Earth System Sciences, 22(2), 1299–1315. https://doi.org/10.5194/hess-22-1299-2018
Donnelly, C., Andersson, J. C., & Arheimer, B. (2016). Using flow signatures and catchment similarities to evaluate the E-HYPE multi-basin model across Europe. Hydrological Sciences Journal, 61(2), 255–273. https://doi.org/10.1080/02626667.2015.1027710
Duan, Q., Sorooshian, S., & Gupta, V. K. (1992). Effective and efficient global optimization for conceptual rainfall-runoff models. Water Resources Research, 28(4), 1015–1031. https://doi.org/10.1029/91WR02985
Duan, Q., Sorooshian, S., & Gupta, V. K. (1994). Optimal use of the SCE-UA global optimization method for calibrating watershed models. Journal of Hydrology, 158(3-4), 265–284. https://doi.org/10.1016/0022-1694(94)90057-4
Eghbal, H. J. (1979). K-S Test for Detecting Changes from Landsat Imagery Data. IEEE Transactions on Systems, Man and Cybernetics, 9(1), 17–23. https://doi.org/10.1109/TSMC.1979.4310069
Li, H., Wigmosta, M. S., Wu, H., Huang, M., Ke, Y., Coleman, A. M., & Leung, L. R. (2013). A physically based runoff routing model for land surface and earth system models. *Hydrology and Earth System Sciences*, 15(5), 2409–2425. https://doi.org/10.1038/hess-19-2409-2015

Guo, J., Liang, X., & Leung, L. R. (2004). A new multiscale flow network generation scheme for land surface models. *Geophysical Research Letters*, 31, L23502. https://doi.org/10.1029/2004GL021381

Hundecha, Y., Arheimer, B., Donnelly, C., & Pechlivanidis, I. (2016). A regional parameter estimation scheme for a pan-European multi-basin model. *Journal of Hydrology: Regional Studies*, 6, 90–111. https://doi.org/10.1016/j.ejrh.2016.04.002

Lehner, B., Verdin, K., & Jarvis, A. (2008). New global hydrography derived from spaceborne elevation data. *Eos, Transactions American Geophysical Union*, 89(10), 93–94. https://doi.org/10.1002/2008EO100001

Leng, G., Huang, M., Voisin, N., Zhang, X., Asrar, G. R., & Leung, L. R. (2016). Emergence of new hydrologic regimes of surface water resources in the conterminous United States under future warming. *Environmental Research Letters*, 11(11), 114003. https://doi.org/10.1088/1748-9326/11/11/114003

Li, H., Wigmosta, M. S., Wu, H., Huang, M., Ke, Y., Coleman, A. M., & Leung, L. R. (2013). A physically based runoff routing model for land surface and earth system models. *Journal of Hydrology*, 143(3), 808–828. https://doi.org/10.1175/JHM-D-12-015.1

Liang, X., Lettenmaier, D. P., Wood, E. F., & Burges, S. J. (1994). A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *Journal of Geophysical Research*, 99(D7), 14,415–14,428. https://doi.org/10.1029/94JD00483

Liang, X., Wood, E. F., & Lettenmaier, D. P. (1996). Surface soil moisture parameterization of the VIC-2L model: Evaluation and modification. *Global and Planetary Change*, 13(1–4), 195–206. https://doi.org/10.1029/95GC00046-1

Lin, P., Pan, M., Beck, H. E., Yang, Y., Yamazaki, D., Frasson, R., et al. (2019). Global Reconstruction of Naturalized River Flows at 2.94 Million Reaches. *Water Resources Research*. https://doi.org/10.1029/2019wr025287

Livneh, B., Rosenberg, E. A., Li, N., Nijssen, B., Andreadis, K. M., et al. (2013). A long term hydrologically based dataset of land surface fluxes and states for the conterminous United States: Update and extensions. *Journal of Climate*, 26(23), 9384–9392. https://doi.org/10.1175/JCLI-D-12-00968.1

Lohmann, D., Mitchell, K. E., Houser, P. R., Wood, E. F., Schaeke, J. C., Robock, A., et al. (2004). Streamflow and water balance inter-comparisons of four land surface models in the North American Land Data Assimilation System project. *Journal of Geophysical Research*, 109, D07S91. https://doi.org/10.1029/2003JD003517

Lohmann, D., Racskhe, E., Nijsen, B., & Lettenmaier, D. (1998). Regional scale hydrology: I. Formulation of the VIC-2L model coupled to a routing model. *Hydrological Sciences Journal*, 43(1), 131–141. https://doi.org/10.1026/026669899492107

Madsen, H., Lawrence, D., Lang, M., Martinkova, M., & Kjeldsen, T. (2014). Review of trend analysis and climate change projections of extreme precipitation and floods in Europe. *Journal of Hydrology*, 519, 3634–3650. https://doi.org/10.1016/j.jhydrol.2014.11.003

Mahat, V., Ramirez, J. A., & Brown, T. C. (2017). Twenty-first-century climate in CMIP5 simulations: Implications for snow and water yield across the contiguous United States. *Journal of Hydrometeorology*, 18(8), 2079–2099. https://doi.org/10.1175/JHM-D-16-0098.1

Maurer, E. P., Wood, A. W., Adam, J. C., Lettenmaier, D. P., & Nijsen, B. (2002). A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States. *Journal of Climate*, 15(22), 3237–3251. https://doi.org/10.1175/1520-0442(2002)015<3237:ALTHBD>2.0.CO;2

McIntyre, N., Lee, H., Wheeler, H., Young, A., & Wagener, T. (2005). Ensemble predictions of runoff in ungauged catchments. *Water Resources Research*, 41, W12434. https://doi.org/10.1029/2005WR004289

Merz, R., Parajka, J., & Bílschel, G. (2011). Time stability of catchment model parameters: Implications for climate impact analyses. *Water Resources Research*, 47, W02531. https://doi.org/10.1029/2010WR009509

Mitchell, K. E., Lohmann, D., Houser, P. R., Wood, E. F., Schaeke, J. C., Robock, A., et al. (2004). The multi-institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system. *Journal of Geophysical Research*, 109, D07S09. https://doi.org/10.1029/2003JD003823

Mizukami, N., Clark, M. P., Gutmann, E. D., Mendoza, P. A., Newman, A. J., Nijsen, B., et al. (2016). Implications of the methodological choices for hydrologic portrayals of climate change over the contiguous United States: statistically downscaled forcing data and hydrologic models. *Journal of Hydrometeorology*, 17(1), 73–98. https://doi.org/10.1175/JHM-D-14-0187.1

Mizukami, N., Clark, M. P., Newman, A. J., Wood, A. W., Gutmann, E. D., Nijsen, B., et al. (2017). Towards seamless large-domain parameter estimation for hydrologic models. *Water Resources Research*, 53, 8020–8040. https://doi.org/10.1002/2016wr020401

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through. Part I. A conceptual models discussion of principles. *Journal of Hydrology*, 10(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6
Water Resources Research

Yaz, B. S., Kao, S.-C., Ashfaq, M., Gao, H., Rastogi, D., & Gangrade, S. (2018). Effects of climate change on streamflow extremes and implications for reservoir inflow in the United States. Journal of Hydrology, 556, 359–370. https://doi.org/10.1016/j.jhydrol.2017.11.027

Naz, B. S., Kao, S.-C., Ashfaq, M., Rastogi, D., Mei, R., & Bowling, L. C. (2016). Regional hydrologic response to climate change in the conterminous United States using high-resolution hydroclimate simulations. Global and Planetary Change, 143, 100–117. https://doi.org/10.1016/j.gloplacha.2016.06.003

Newman, A. J., Mizukami, N., Clark, M. P., Wood, A. W., Nijssen, B., & Nearing, G. (2017). Benchmarking of a physically based hydrologic model. Journal of Hydrometeorology, 18(8), 2215–2225. https://doi.org/10.1175/JHM-D-16-0264.1

Nijssen, B., O’Donnell, G. M., Lettenmaier, D. P., Lohmann, D., & Wood, E. F. (2001). Predicting the discharge of global rivers. Journal of Climate, 14(15), 3307–3323. https://doi.org/10.1175/1520-0442(2001)014<3307:PTDOGR>2.0.CO;2

Ouibeidillah, A. A., Kao, S.-C., Ashfaq, M., Naz, B. S., & Tootle, G. (2014). A large-scale, high-resolution hydrological model parameter data set for climate change impact assessment for the conterminous US. Hydrology and Earth System Sciences, 18(1), 67–84. https://doi.org/10.1016/j.hess.18-67-2014

Oudin, L., Andréassian, V., Perrin, C., Michel, C., & Le Moine, N. (2008). Spatial proximity, physical similarity, regression and ungauged catchments: A comparison of regionalization approaches based on 913 French catchments. Water Resources Research, 44, W03413. https://doi.org/10.1029/2007WR006240

Pan, M., Sheffield, J., Wood, E. F., Mitchell, K. E., Houser, P. R., Schaake, J. C., et al. (2003). Snow process modeling in the North American Land Data Assimilation System (NLDAS): 1. Evaluation of model simulated snow water equivalent. Hydrology and Earth System Sciences, 7(3), 157–171. https://doi.org/10.5194/hess-7-157-2003

Ray, R. L., Beighley, R. E., & Yoon, Y. (2016). Integrating runoff generation and flow routing in Susquehanna River Basin to characterize key hydrologic processes contributing to maximum annual flood events. Journal of Hydrologic Engineering, 21(9), 04016026. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001389

Samaniego, L., Kumar, R., & Attinger, S. (2010). Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. Water Resources Research, 46, W05523. https://doi.org/10.1029/2008WR007327

Schumann, G. P., Neal, J. C., Voisin, N., Andreadis, K. M., Pappenberger, F., Phanthurongpakdee, N., et al. (2013). A first large-scale flood inundation forecasting model. Water Resources Research, 49, 6248–6257. https://doi.org/10.1002/wrcr.20521

Seyyedi, H., Anagnostou, E. N., Beighley, E., & McCollum, J. (2014). Satellite-driven downscaling of global reanalysis precipitation products for hydrological applications. Hydrology and Earth System Sciences, 18(12), 5077–5091. https://doi.org/10.5194/hess-18-5077-2014

Seyyedi, H., Anagnostou, E. N., Beighley, E., & McCollum, J. (2015). Hydrologic evaluation of satellite and reanalysis precipitation datasets over a mid-latitude basin. Atmospheric Research, 164-165, 37–48. https://doi.org/10.1016/j.atmosres.2015.03.019

Shafii, M., & Tolson, B. A. (2015). Optimizing hydrologic consistency by incorporating hydrologic signatures into model calibration objectives. Water Resources Research, 51, 3796–3814. https://doi.org/10.1002/2014WR016520

Sheffield, J., Pan, M., Wood, E. F., Mitchell, K. E., Houser, P. R., Schaake, J. C., et al. (2003). Snow process modeling in the North American Land Data Assimilation System (NLDAS): 1. Evaluation of model-simulated snow cover extent. Journal of Geophysical Research, 108(D22), 8849. https://doi.org/10.1029/2002JD003274

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. The Annals of Mathematical Statistics, 19(2), 279–281. https://doi.org/10.1214/aoms/1177730256

Troy, T. J., Wood, E. F., & Sheffield, J. (2008). An efficient calibration method for continental-scale land surface modeling. Water Resources Research, 44, W09411. https://doi.org/10.1029/2007WR006513

Wagen, T., & Gupta, H. V. (2005). Model identification for hydrological forecasting under uncertainty. Stochastic Environmental Research and Risk Assessment, 19(6), 378–387. https://doi.org/10.1007/s00477-005-0077-6

Weilland, F. C. S., Vrugt, J. A., Weerts, A. H., & Bierkens, M. F. (2015). Significant uncertainty in global scale hydrological modeling from precipitation data errors. Journal of Hydrology, 529, 1095–1115. https://doi.org/10.1016/j.jhydrol.2015.08.061

Willner, S. N., Levermann, A., Zhao, F., & Frieler, K. (2018). Adaptation required to preserve future high-end river flood risk at present levels. Science advances, 4(11), eaao1914. https://doi.org/10.1126/sciadv.aao1914

Winsemius, H., Schaeffli, B., Montanari, A., & Savenije, H. (2009). On the calibration of hydrological models in ungauged basins: A framework for integrating hard and soft hydrological information. Water Resources Research, 45, W12422. https://doi.org/10.1029/2009WR007796

Wu, H., Adler, R. F., Tian, Y., Huffman, G. J., Li, H., & Wang, J. (2014). Real-time global flood estimation using satellite-based precipitation and a coupled land surface and routing model. Water Resources Research, 50, 2693–2717. https://doi.org/10.1002/2013WR014710

Wu, H., Gu, G., Yan, Y., Gao, Z., & Adler, R. F. (2018). Global Flood Monitoring Using Satellite Precipitation and Hydrological Modeling. In G. J.-P. Schumm, P. D. Bates, H. Apel, & G. T. Aronica (Eds.), Global Flood Hazard: Applications in Modeling, Mapping, and Forecasting (pp. 87–113). Hoboken, NJ and Washington, DC: John Wiley & Sons and American Geophysical Union.

Xia, Y., Mitchell, K., Ek, M., Cosgrove, B., Sheffield, J., Lai, L., et al. (2012). Continental-scale water and energy flux analysis and validation for North American Land Data Assimilation System project phase 2 (NLDAS-2): 2. Validation of model-simulated streamflow. Journal of Geophysical Research, 117, D03110. https://doi.org/10.1029/2011JD016051

Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E. F., et al. (2012). Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2). 1. Intercomparison and application of model products. Journal of Geophysical Research, 117, D03109. https://doi.org/10.1029/2011JD016048

Xia, Y., Mocko, D. M., Wang, S., Pan, M., Kumar, S. V., Petes-Lidard, C. D., et al. (2018). Comprehensive evaluation of the variable infiltration capacity (VIC) model in the North American Land Data Assimilation System. Journal of Hydrometeorology, 19(11), 1853–1879. https://doi.org/10.1175/JHM-D-18-0139.1

Yadav, M., Wagen, T., & Gupta, H. (2007). Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins. Advances in Water Resources, 30(8), 1756–1774. https://doi.org/10.1016/j.advwatres.2007.01.005

Yamazaki, D., Almeida, G. A., & Bates, P. D. (2013). Improving computational efficiency in global river models by implementing the local inertial flow equation and a vector-based river network map. Water Resources Research, 49, 7221–7235. https://doi.org/10.1002/wrcr.20552
Yamazaki, D., Kanae, S., Kim, H., & Oki, T. (2011). A physically based description of floodplain inundation dynamics in a global river routing model. *Water Resources Research, 47*, W04501. https://doi.org/10.1029/2010WR009726

Yilmaz, K. K., Gupta, H. V., & Wagener, T. (2008). A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model. *Water Resources Research, 44*, W09417. https://doi.org/10.1029/2007WR006716

Yin, J., Lin, N., & Yu, D. (2016). Coupled modeling of storm surge and coastal inundation: A case study in New York City during Hurricane Sandy. *Water Resources Research, 52*, 8685–8699. https://doi.org/10.1002/2015WR019102

Yin, J., Yu, D., Lin, N., & Wilby, R. L. (2017). Evaluating the cascading impacts of sea level rise and coastal flooding on emergency response spatial accessibility in Lower Manhattan, New York City. *Journal of Hydrology, 555*, 648–658. https://doi.org/10.1016/j.jhydrol.2017.10.067

Yu, D., & Lane, S. N. (2006). Urban fluvial flood modelling using a two-dimensional diffusion-wave treatment, part 1: mesh resolution effects. *Hydrological Processes: An International Journal, 20*(7), 1541–1565. https://doi.org/10.1002/hyp.5935

Zhang, X.-J., Tang, Q., Pan, M., & Tang, Y. (2014). A long-term land surface hydrologic fluxes and states dataset for China. *Journal of Hydrometeorology, 15*(5), 2067–2084. https://doi.org/10.1175/JHM-D-13-0170.1

Zhao, Y., & Beighley, E. (2016). Upscaling Surface Runoff Routing Processes in Large-Scale Hydrologic Models: Application to the Ohio River Basin. *Journal of Hydrologic Engineering, 22*(4), 04016068. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001478