ABSTRACT: This study assesses a large-scale hydrologic modeling framework (WRF-Hydro-RAPID) in terms of its high-resolution simulation of evapotranspiration (ET) and streamflow over Texas (drainage area: 464,135 km²). The reference observations used include eight-day ET data from MODIS and FLUXNET, and daily river discharge data from 271 U.S. Geological Survey gauges located across a climate gradient. A recursive digital filter is applied to decompose the river discharge into surface runoff and base flow for comparison with the model counterparts. While the routing component of the model is pre-calibrated, the land component is uncalibrated. Results show the model performance for ET and runoff is aridity-dependent. ET is better predicted in a wet year than in a dry year. Streamflow is better predicted in wet regions with the highest efficiency ~0.7. In comparison, streamflow is most poorly predicted in dry regions with a large positive bias. Modeled ET bias is more strongly correlated with the base flow bias than surface runoff bias. These results complement previous evaluations by incorporating more spatial details. They also help identify potential processes for future model improvements. Indeed, improving the dry region streamflow simulation would require synergistic enhancements of ET, soil moisture and groundwater parameterizations in the current model configuration. Our assessments are important preliminary steps towards accurate large-scale hydrologic forecasts.

(KEY TERMS: evapotranspiration; streamflow; surface runoff; base flow; MODIS; WRF-Hydro; Noah-MP; RAPID.)

INTRODUCTION

The ability and accuracy of weather prediction have greatly improved over the past decade, resulting in transformation of hydrologic forecasting approaches from statistical and stochastic methods based on recorded observations (McEnery et al., 2005) to deterministic and probabilistic forecasts based on numerical weather prediction (NWP) tools.
(Cloke and Pappenberger, 2009). Numerous studies have highlighted recent efforts in developing such forecasting systems for Europe (Thiemig et al., 2009), Africa (Thiemig et al., 2010), South America (Paiva et al., 2011), and the entire globe (Pappenberger et al., 2012; Alfieri et al., 2013; Winsemius et al., 2013).

Likewise, a continental-scale high-resolution hydrologic forecasting system for the United States (U.S.) has been developed recently (Maidment, 2016), which literally transforms a real-time weather map to a river forecast map, using an earth system modeling framework. This hydrologic forecasting system uses the community WRF-Hydro model (Gochis et al., 2015), an architectural framework which leverages the U.S. operational weather forecasting, the state-of-the-art land surface models (LSMs), and the geo-referenced vector river network — the National Hydrography Dataset Plus (NHDPlus), providing streamflow forecasts at 2.67 million river reaches operationally across the contiguous U.S. Comparing to previous gauge-based forecasting points at ~4,000 nationally (McEnery et al., 2005), this new forecasting system represents a factor of 741 increase in spatial coverage, signaling a substantial shift from watershed hydrology to continental hydrology.

The National Flood Interoperability Experiment (NFIE), a research collaboration among academia, federal agencies, and private partners, was proposed to translate the unprecedented information on river flow forecasts to inundation maps, which will eventually inform local emergency response planning (Maidment, 2016). Since the onset of NFIE, continuous efforts have been made in a wide range of aspects, including the development of baseline geospatial hydrologic data for the continental flood forecasting system (Viger et al., 2016), the river hydraulics and inundation mapping capability at high resolution (Follum et al., 2016), and the development of web platforms for warning delivery to emergency responders and the general public (Swain, 2015). While these efforts are related mostly to the communication and the translation of hydrologic simulations, their potential strengths and weaknesses in terms of water balances have not been well evaluated. Such evaluation is essential for ensuring the robustness of the hydrologic modeling framework across the intended spatiotemporal scales.

As the demonstration version of the continental-scale river modeling system, the WRF-Hydro-RAPID framework was used at the 2015 NFIE Summer Institute (Maidment, 2016). Within this framework, the refined Noah LSM with multi-parameterization options (hereafter Noah-MP) (Niu et al., 2011) and the Routing Application for Parallel computation of Discharge (RAPID) (David et al., 2011) were used for hydrologic simulations. While global- and continental-scale evaluations of Noah-MP have been documented (e.g., Yang et al., 2011; Cai et al., 2014a), these studies implemented the model at a resolution usually coarser than 0.125° and used limited gauge locations at major rivers and basin outlets as a reference for model evaluation. David et al. (2013) evaluated RAPID’s performance using runoff outputs from multiple LSMs, with a focus on the routing parameter optimization, but they did not establish linkages between streamflow prediction errors with biases in LSM-simulated water fluxes. Therefore, it is important to assess the model performance at finer spatiotemporal scales than previous studies and establish the relationships between the errors in the modeled streamflow and the biases in runoff components.

This study uses WRF-Hydro-RAPID over Texas in the offline mode such that Noah-MP is run in a stand-alone mode (i.e., uncoupled to WRF) to simulate land surface hydrology (Gochis et al., 2015). Due to emerging interests in using satellite-based evapotranspiration (ET) data to calibrate hydrologic models (e.g., Kunnath-Poovakka et al., 2016), this study also evaluates the modeled ET against the Moderate Resolution Imaging Spectroradiometer (MODIS) ET to understand the potential usefulness of the remotely sensed ET product in improving streamflow predictions.

While the current operational U.S. National Water Model (NWM) has evolved from the NFIE initiative, it has several advanced features compared to the model configurations presented in this study. Specifically, NWM has a streamflow data assimilation capability, a more physically based hydrologic routing scheme, a feature allowing simple level-pool routing for over 1,200 major reservoirs, and tools for model calibrations (http://water.noaa.gov/about/nwm).

The aim of this study is not to obtain the best-achievable prediction skills as required in an operational setting. Instead, it is to establish a better understanding of the model behaviors across a climate gradient with a focus on the interplay between biases in ET and individual runoff components. Based upon the default Noah-MP configurations, the results presented here should be viewed as a reference for the baseline model performance, with a hope to benefit future model users who work on specific watersheds in Texas and similar regions and to guide future model improvement activities.

The model and data used in this study are first described, followed by the methodology. Next, results and discussions on assessing the modeled ET, daily streamflow, and the individual runoff components are presented. The last section summarizes the major conclusions of this study.
MODEL AND DATA

Model Description and Experimental Configurations

WRF-Hydro (Gochis et al., 2015) is the hydrological extension of the Weather Research and Forecasting model (WRF), a next-generation NWP tool for research and operational purposes, which has a large worldwide user community mainly in atmospheric science (http://www.wrf-model.org/index.php). As a model coupling architecture for easier linkage between hydrology simulations with atmospheric models, WRF-Hydro incorporates horizontal routing processes and water resource management and provides several LSM options and hydrologic routing options to serve various hydrometeorological applications. Among the several routing options, WRF-Hydro supports the alternatives of grid-based and reach-based schemes (Gochis et al., 2015), one of which is the RAPID model. RAPID directly resolves streamflow computations on the NHDPlus river network in a vector-based environment (David et al., 2011). Compared to grid-based routing models, vector-based models have the advantages of higher computational efficiency, and more effective hazard communication and visualization due to its geographic information system (GIS) capabilities (Lehner and Grill, 2013; Yamazaki et al., 2013), and thus WRF-Hydro-RAPID is a hybrid framework consisting of both grid- and vector-based modeling units. WRF-Hydro encapsulates several state-of-the-art LSMs including Noah-MP (Niu et al., 2011); Noah-MP is used in this study because it was adopted by both the NFIE and the operational NWM to simulate land surface hydrology. The model is configured at ~5 km × 5 km grid resolution (269 × 314 grid cells in total) and driven by 3-hourly atmospheric forcing from the North American Land Data Assimilation (NLDAS-2). The simulation period is from January 1, 2008 to December 31, 2011. Noah-MP provides multiple physical parameterization options within one model, from which the runoff parameterization scheme with a simple 1-dimensional groundwater model (Niu et al., 2007) is used. Cai et al. (2014a) found that the groundwater model in Noah-MP requires 34 years for spinning up for the Mississippi River Basin. Texas is drier than the Mississippi as a whole, so we conduct a 100-year spin-up process for our domain by driving the model with the 2008 atmospheric forcing for 100 times. The gridded total runoff is passed to RAPID, using a catchment centroid-based grid-to-vector coupling interface (Lin et al., 2015) to simulate horizontal river transport. RAPID uses a matrix-based Muskingum method to solve horizontal routing, and the Muskingum parameters k (travel time) and x (weighting factor) are obtained from an optimal set derived by David et al. (2013). Noah-MP is kept uncalibrated with its default parameter values as suggested by Niu et al. (2011) and Cai et al. (2014a). The time steps of Noah-MP and RAPID are 3-h and 15-min, respectively; the model outputs are averaged to daily values from 2008 to 2011.

Observational Data

The 15-min river discharge data are obtained from the U.S. Geological Survey (USGS) gauge stations in the Texas Hydrologic Region 12 using the HydroDesktop batch downloading tool (Ames et al., 2012). In total, the Region 12 contains 731 USGS gauges, but only 271 of them have continuous time series covering the entire study period (2008-2011). Since filling out missing values would artificially introduce additional uncertainties, we only use these 271 gauge observations in the model evaluation. We use the remote sensing-derived eight-day actual ET data from the MODIS MOD 16 product (Mu et al., 2011). Despite the uncertainties commonly seen in remotely-sensed estimates, the MOD16 ET is used as a reference to compare with the modeled ET because it is arguably one of the best available large-scale ET datasets with a reasonably high spatiotemporal resolution and well-documented uncertainty assessments. Mu et al. (2013) found that the MODIS ET product has ~0.3 mm/day of average mean absolute error with respect to 46 FLUXNET site estimates across North and South America. An independent site-specific assessment showed that the root-mean-square-error of the MOD16 product ranged from 26 to 32 mm per month (Velpuri et al., 2013). In this study, MODIS eight-day total ET data at ~1-km grid spacing are geo-referenced, spatially re-scaled, and aggregated using an automatic data processor tool (Rajib et al., 2016) to match the ~5-km model grids over Texas (three MODIS tiles covering the main portions of Texas are used). As an additional reference, the in situ ET measurement from a FLUXNET station (Freeman Ranch-Woodland, Texas, 29.9495°N, 97.9962°W) is also used to compare with both MODIS ET and model-simulated ET. This FLUXNET station is the only one of the four available FLUXNET sites within our study domain that contains sufficient time-series data for the simulation period. All the above-mentioned observational datasets are used only for model evaluation, but not in any form of parameter estimation/calibration purposes.

Study Domain

Figure 1 shows the study domain, the major river basins, and the distribution of climatological aridity index (AI) of the region. The LSM domain covers the
entire state of Texas to represent a strong west-east precipitation gradient, while the river routing domain is set up for the Hydrologic Region 12, which drains an area of 464,135 km² with 68,143 NHDPlus river reaches excluding the Rio Grande, the Canadian, and the Red River Basins based on the NHDPlus domain discretization. Defined as the ratio of annual precipitation to annual potential evapotranspiration (PET), AI is calculated, using the 36-year average NLDAS-2 monthly forcing data. A high (low) AI value indicates a humid (arid) climate. According to the climate types defined by the United Nations Environmental Programme, Texas has an arid climate (AI < 0.2) in the west, a semiarid climate (0.2 ≤ AI ≤ 0.5) in the central, and a subhumid to humid climate (AI > 0.5) in the east. In this paper, we mainly refer to the dry (wet) region of Texas as regions with an AI ≤ 0.3 (vs. AI > 0.3) to reflect the west-east climatic contrast. The state has a large precipitation gradient with the annual rainfall ranging from 200 to 1,400 mm, and it has high vulnerability to extreme events. The Gulf Coast is threatened by frequent hurricane-induced flooding, while central/western Texas experiences recurrent flash droughts and floods.

**METHODOLOGY**

**Runoff Separation**

A semi-empirical two-parameter recursive digital filter (Eckhardt, 2005, 2008) is used to partition the streamflow data into surface runoff and base flow components. This technique relies on the notion that high frequency variability of the streamflow is primarily caused by direct runoff. Hence, base flow can be identified by low-pass filtering of the streamflow hydrograph. Within its semi-automated structure, the filter algorithm takes the streamflow hydrograph, an exponential base flow recession constant, $a$, and the maximum value of base flow index (BFI$_{\text{max}}$, long-term ratio of base flow to total streamflow) as the user inputs. While $a$ can be determined by recession analyses (e.g., Tallaksen, 1995; Sujono et al., 2004), BFI$_{\text{max}}$ is a nonmeasurable quantity and a fitting constant for the algorithm (Eckhardt, 2005).

In this study, daily streamflow hydrographs from all the 271 USGS gauges are filtered to obtain corresponding daily time series of surface runoff and base flow. In a separate setting before inserting into the filter algorithm, each of the streamflow time series covering their respective maximum range of available data until 2014 is analyzed to identify the dates showing zero streamflow. Catchments having no streamflow at the outlet for more than 50% of the days within a year and consistently for each year of their entire record are characterized as ephemeral. As BFI$_{\text{max}}$ cannot be determined prior to separation, two typical values are used over the entire study domain depending on the perennial (BFI$_{\text{max}}$ = 0.8) and ephemeral (BFI$_{\text{max}}$ = 0.5) nature of the catchments (e.g., Eckhardt, 2005; Partington et al., 2012). Since it is inefficient to conduct numerical recession analysis for calculating $a$ for each of the streamflow hydrographs, a constant value of 0.96 is adopted from a multi-model comparison conducted by Eckhardt (2008) on 65 North American watersheds. To tentatively assess the sensitivity of these chosen values, BFI$_{\text{max}}$ and $a$ are varied one at a time by ±0.20 and ±0.03 for four randomly chosen gauge stations, which produces a maximum of 4% and 9.5% difference in the average annual surface runoff, respectively (not shown here). Such small differences in response to the variations in BFI$_{\text{max}}$ and $a$ indicate the representativeness of these chosen values, although $a$ shows a relatively stronger influence on the filter results compared to BFI$_{\text{max}}$ (Eckhardt, 2012). The overall performance of the filter is compared against a recent study by the USGS and the Texas Water Development Board (Baldys and Schalla, 2016). Table 1 shows that the average BFI calculated, using Eckhardt’s filter for five representative USGS gauge stations in the Brazos River Basin is only 8% higher than that reported by Baldys and Schalla (2016), using the local-minimum method of Hydrograph Separation and Analysis (HYSEP) program.

The same filter is also applied to the model simulated streamflow at locations of the 271 USGS
TABLE 1. Comparison of the Calculated Eckhardt’s BFI with the HYSEP-Derived BFI by Baldys and Schalla (2016).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>08082000</td>
<td>0.2</td>
<td>0.18</td>
</tr>
<tr>
<td>08080500</td>
<td>0.2</td>
<td>0.16</td>
</tr>
<tr>
<td>08084000</td>
<td>0.35</td>
<td>0.32</td>
</tr>
<tr>
<td>08085500</td>
<td>0.24</td>
<td>0.24</td>
</tr>
<tr>
<td>08085000</td>
<td>0.4</td>
<td>0.32</td>
</tr>
<tr>
<td>Average</td>
<td>0.27</td>
<td>0.24</td>
</tr>
</tbody>
</table>

Note: BFI, base flow index; HYSEP, Hydrograph Separation and Analysis program with local minimum method; USGS, U.S. Geological Survey.

Skill Metrics for Daily Streamflow Evaluation

Four skill metrics are used to evaluate the model performance in simulating daily streamflow, which includes Pearson correlation coefficient (CC), Nash-Sutcliffe efficiency (NSE), normalized root-mean-square-error (NRMSE), and normalized bias (NBIAS). In Equations (1-4), $Q_{\text{obs}}$ and $Q_{\text{mod}}$ denote observed and model-simulated streamflow, respectively; $N$ is the number of days. While these skill metrics are mathematically related (Gupta et al., 2009), each of them is interpreted with a slightly different focus to facilitate a more comprehensive understanding on the model performance. CC (Equation 1) measures the proportion of the total variance of the observed data that can be explained by the model (Legates and McCabe, 1999), which cannot be used to assess model biases. NSE (Equation 2) and RMSE, as the two most commonly used criteria in hydrological evaluations, measure both the variability of time series and the magnitude of errors (Gupta et al., 2009). NRMSE normalizes the squared model error, using the variance of the observed data, and therefore, to provide an objective assessment of the model error across spatial scales without mixing up with large/small streams, NRMSE (Equation 3) is used. NBIAS (Equation 4) is calculated to identify both the magnitude and the sign of errors in the mean streamflow.

\[
CC = \frac{\text{cov}(Q_{\text{obs}}, Q_{\text{mod}})}{\sigma_{\text{obs}} \sigma_{\text{mod}}} \tag{1}
\]

\[
\text{NSE} = 1 - \frac{\sum_{i=1}^{N} (Q_{\text{mod}} - Q_{\text{obs}})^2}{\sum_{i=1}^{N} (Q_{\text{obs}}^i - \bar{Q}_{\text{obs}})^2} \tag{2}
\]

\[
\text{NRMSE} = \sqrt{\frac{\sum_{i=1}^{N} (Q_{\text{mod}}^i - Q_{\text{obs}}^i)^2}{\sum_{i=1}^{N} (Q_{\text{obs}}^i)^2}} \tag{3}
\]

\[
\text{NBIAS} = \frac{\sum_{i=1}^{N} (Q_{\text{mod}}^i - Q_{\text{obs}}^i)}{\sum_{i=1}^{N} Q_{\text{obs}}^i} \tag{4}
\]

RESULTS AND DISCUSSION

Evapotranspiration

Figure 2 shows the ET difference between the model and the MODIS ET averaged across the four seasons of a wet year (2008) and a dry year (2011).

Regardless of wet/dry years or seasons, it is seen that the model consistently underpredicts ET over the wet southeastern Texas and overpredicts over the drier northwestern Texas and the Mexican Plateau region when compared against MODIS ET data. The magnitude of the difference represents 0-50% underprediction in the wet region and more than 100% overprediction in the arid region, respectively. Difference between modeled and MODIS ET is the largest in the summer (JJA) and smallest in the winter (DJF); in DJF the difference is generally less than 0.3 mm/day, which is comparable to the documented uncertainty in the MODIS ET estimates (e.g., Mu et al., 2013). By comparing the spatially averaged monthly MODIS ET with four NLDAS-2 LSMs, Long et al. (2014) observed a more dampened seasonal cycle in remote sensing-derived ET (hereafter RS-ET) products. This was explained by the limited responsiveness to precipitation in RS-ET products mostly due to a lack of soil wetness constraints, which is explicitly accounted for in the LSM-ET estimates. Compared to the generic finding on the “model overestimation” problem in JJA by Long et al. (2014), Figure 2 detects spatially explicit details on model behavior showing both overprediction and underprediction tendency of the model during the same season. Such a spatial pattern of the ET over- and underpredictions across Texas can also be seen in three other NLDAS-2 LSMs (Cai et al., 2014b),
suggesting possible model structural or parameterization problems in simulating ET over contrasting climatic regions.

Although both the MODIS ET and the LSM ET use the Penman-Monteith (PM) approach, the main differences between the two estimates can be attributed to three factors: (1) the driving meteorology and energy data, or the forcing data difference, (2) the treatment of vegetation (i.e., whether the leaf area index (LAI) is prescribed or predicted; if prescribed, whether it is based on long-term climatology or satellite data for specific years), and (3) the treatment of soil wetness constraints. The spatial difference in ET is negatively correlated ($p < 0.05$) with aridity (Figure 3), an integrated climatic factor which implicitly encompasses the land heterogeneity including soil, vegetation, and geological conditions. Therefore, as detailed below, it is necessary to clarify several land surface physical parameters and resistance terms that are directly used in the PM approach, which would help shed light on the distinct spatial pattern of the ET flux difference in our case.

In the current version of Noah-MP, soil surface resistance ($r_s$), a term describing the restriction for water vapor to diffuse from soil surface into the air, uses a parameterization proposed by Sakaguchi and Zeng (2009) (hereafter SZ2009) being updated from the previous parameterization of Sellers et al. (1992). In SZ2009, $r_s$ is defined to account for soil wetness and texture factors as shown in Equations (5-7).

$$r_s = \frac{L}{D}$$

$$D = D_0 \theta_{sat}^2 \left(1 - \frac{\theta_r}{\theta_{sat}}\right)^{2+3b}$$

$$L = \frac{d_1 \exp[(1 - \theta_1/\theta_{sat})^w] - 1}{e^{-1}},$$

where $D$ represents vapor diffusivity within the soil (m$^2$/s), which is a function of soil texture, saturation soil moisture $\theta_{sat}$, and residual water content $\theta_r$; and $L$ represents dry layer thickness ($m$), which is a
function of top layer soil wetness $\theta_1/\theta_{sat}$. There are two parameters ($b$ and $w$) and three constants ($D_0$, $d_1$, and $e$) in the SZ2009 parameterization: $b$ is the soil texture parameter (Clapp and Hornberger, 1978); $w$ is an empirical parameter controlling the exponential relationship between $L$ and $\theta_1/\theta_{sat}$. $D_0$ is the molecular diffusion coefficient of water vapor in the air ($2.2 \times 10^{-5}$ m$^2$/s); $d_1$ is the thickness of topsoil layer (0.1 m); and $e$ is 2.718. SZ2009 improves the previous soil evaporation parameterization by better constraining soil wetness (Sakaguchi and Zeng, 2009), yet the parameter controlling how $L$ varies with the top layer soil wetness (i.e., $w$ in Equation 7) is set to 5 globally based on limited empirical field experiments. It is hence likely that the large wet/dry climate gradient may not be well-represented, using a constant value for $w$ in spatially distributed simulations, given the sensitivity of $r_e$ to $w$ has been documented (Sakaguchi and Zeng, 2009) and that of ET to $w$ is observed in our preliminary investigation (not shown here). By testing 139 hard-coded parameters in Noah-MP, a recent study (Cuntz et al., 2016) also concluded that $w$ is the most sensitive yet overlooked parameter for ET and runoff simulations. Therefore, it is suggested for future studies to adjust or derive spatially distributed values for parameter based on RS-ET data, or to revisit the soil surface resistance parameterization in Noah-MP by developing more appropriate process-based schemes (e.g., Zeng et al., 2011; Zhang et al., 2015) in order to dampen the model’s oversensitivity to $w$ (Cuntz et al., 2016). These could potentially improve the overall ET simulation, especially in western Texas where the direct soil evaporation is a major component due to the relatively sparse vegetation on land surface. For the wet southeastern Texas where canopy transpiration plays an important role, treatment of stomatal resistance ($r_{stomatal}$) (Ball et al., 1987), LAI (which defines structural parameter for the vegetation greenness), and mechanisms such as the root water uptake (Teuling et al., 2006) could also lead to ET biases. Due to a lack of observations for these individual components, however, this study falls short in concluding which transpiration estimate (model or MODIS) is more reliable. Accordingly, a more comprehensive understanding on the individual ET components (i.e., soil evaporation and plant transpiration) should benefit from finer resolution ET measurements based on much advanced algorithms in near future to better inform hydrologic modeling.

Prominent ET differences between model and MODIS are also found over two noticeable blue bands in the northeast-southwest direction of the southeastern Texas (Figure 2), where model predicts a much smaller ET flux than the MODIS. This is a region that roughly runs 480 km from the Red River in North Texas to San Antonio in the south, where the input soil texture is primarily sand. Among the 12 soil categories used in the model, sand supposedly holds the least available water to sustain ET, and thus the model predicts a smaller ET over the region compared to the MODIS ET, an estimate which does not explicitly consider the soil texture control on ET fluxes.

In Figure 4, we plot the CC between the model and the MODIS eight-day ET time series for 2008 and 2011. At each ~5-km grid cell, CC is calculated using 48 data samples along the whole year to indicate the agreement in temporal variability between the model and the MODIS ET estimates. The two estimates have a better temporal agreement in a normal to wet year (2008) than an extreme drought year (2011), as indicated by more grid cells with a relatively higher CC in Figure 4a than Figure 4b. 2011 was an extreme drought year for Texas, and a large extent of Texas observed reductions in greenness (Sun et al., 2015). In our model configuration, however, a static monthly LAI based on the MODIS LAI climatology is used (the default Noah-MP LAI option). This may result in incorrect greenness representations in the model, and consequently, erroneous hydrologic simulations. Noah-MP also provides a dynamic vegetation (DV) option, where vegetation-related variables are calculated based on the climate, soil, and vegetation conditions at each model time step, albeit with its uncertainties and limitations. The results thus suggest the need for utilizing the DV option in real-time hydrologic forecasting cases and carefully evaluating the coupled vegetation-hydrologic processes, which could improve the temporal water fluxes simulation especially during extreme wet or dry conditions.

Figure 5 shows the time series of model-simulated ET with corresponding MODIS estimates in 2008 and 2011 at four different locations. One location is a FLUXNET site (Freeman Ranch, Woodland) in which case we assume that the in situ measurement (point data) is representative of the average hydrologic condition in the particular 5-km grid cell where the tower is located. The results are consistent with Figure 2 as the model and MODIS estimates have larger discrepancies in JJA than in other seasons. Besides, the model seems to overpredict ET fluxes in the dry region compared to MODIS. At Locations A and B, model and MODIS ET are in good agreement in both years; at Location C, MODIS ET is able to show a drop in ET fluxes in a drought year while the model still predicts falsely high ET fluxes in that year. From all locations, it is apparent that modeled ET tends to show more responsiveness to high ET values temporally, which is consistent with Long et al. (2014). At the FLUXNET site, MODIS ET shows a
relatively better match with the *in situ* measurements in terms of mean ET values, especially for 2008 where the mean MODIS ET is ~0.3 mm/day closer to the *in situ* measurements than modeled ET. On the other hand, the model tends to imitate the temporal variability of the *in situ* measurements better, especially in capturing the higher ET peaks during both 2008 and 2011. Note that the site has an aridity of ~0.4 and the land cover type is woody savannah whereas the model’s dominant land cover
The type is cropland/grassland mosaic. The closer match between MODIS ET and site observations in mean values but not in temporal variability may justify the use of MODIS ET as a spatial reference dataset to adjust model simulations, although its applicability as a temporal reference especially in regions of diverse aridity should be comprehensively investigated in the future.

Streamflow

Figure 6 presents the model performance in simulating daily streamflow at 271 USGS gauges based on a priori parameter set of the default WRF-Hydro-RAPID configuration as described in the Model and Data section. Complementary to previous studies evaluating Noah-MP at the outlets of major river basins (Yang et al., 2011; Cai et al., 2014a), our evaluation includes all the available gauges to provide a more comprehensive perspective of the model performance that has more practical values to the end users of hydrologic forecasts. Since the routing component of this modeling framework performs streamflow computations at the level of individual NHDPlus river reaches, each streamflow prediction is associated with one exact geographic location and a unique identification, making it easier to evaluate the simulated streamflow against hundreds of USGS gauge observations (Figure 6).

Four skill metrics for the daily streamflow simulation are spatially mapped in Figure 6. It is seen that central and northeastern Texas has relatively higher CC and NSE values (Figures 6a and 6b), suggesting reasonably good model performances over the subhumid to humid regions of Texas, especially in the upper Trinity, the lower Colorado, and the San Jacinto River Basins where the NSE generally exceeds 0.5. The best model performance is seen at the Onion Creek at Highway 183 (USGS gauge # 08159000), with the daily NSE equaling 0.66 based on the uncalibrated modeling results. With a drainage basin of 831.4 km², this gauge was one of the most severely flooded areas in central Texas during the 2013 Halloween flood, and the results clearly highlight the model’s potential in issuing effective hydrologic forecasts for this particular gauge (or the adjacent regions in general). However, large biases are seen in the western part and the Nueces River Basin, as displayed by negative NSE (Figure 6b) and large NRMSE and NBIAS values (Figures 6c and 6d) over these regions. Calibrating the LSM component (Noah-MP) might have produced a much better fit with the observed streamflow hydrographs, but would impart an unknown degree of undetectable bias in the surface runoff and base flow fluxes because of

![Figure 6](image_url)
parameter equifinality (Beven and Freer, 2001). Thus, it is the goal of this study to objectively document the performance of the default WRF-Hydro-RAPID modeling framework, instead of conducting model calibration.

Figure 7 shows daily streamflow hydrographs at four USGS gauges, along with their performance metrics against respective observed data. Figures 7a-7c represent three gauges which have relatively good model performances with the NSEs equal to 0.66, 0.56, and 0.46, respectively. This is an indicator of the model being quite effective in these areas with its uncalibrated configuration. On the contrary, Figure 7d represents a gauge located in the drier western Texas showing a poor NSE, which is mainly due to the overpredicted base flow and the flashiness in surface runoff in the model simulation. In that part of Texas, the NSE values based on monthly averaged streamflow are not better than the daily NSE values (Figures 7e and 7f), suggesting a persistent problem in the LSM to generate the correct amount of total runoff, which is not an issue likely to be resolved from better physical representations of the horizontal flow routing processes. This is in line with the previous studies (Xia et al., 2012; Cai et al., 2014b) assessing monthly/annual streamflow simulation skills by the NLDAS-2 LSMs in 961 small watersheds across the U.S. Similar to our findings, Cai et al. (2014a, b) and Xia et al. (2012) indicated relatively larger biases in simulated streamflow, especially across the drier Great Plains. However, all the previous studies included a very limited portion of north/northwestern Texas, not to mention their coarser timescale. The distinct spatial pattern of streamflow bias as shown in Figures 6 and 7 resonates and complements the general findings from previous studies regarding a potential weakness of the LSM in capturing semiarid to arid region processes, suggesting the need for future model improvements or applying bias corrections over these regions to fulfill operational goals.

**Surface Runoff and Base Flow Components**

To further understand the error sources in the river flow predictions, Figure 8 provides a large-scale evaluation of the LSM-simulated surface runoff and base flow components. It is clearly seen that surface runoff is underpredicted in the humid eastern Texas...
(Figure 8a, warm colors), whereas base flow is over-predicted by the model (Figure 8b, cold colors). The magnitude of difference between model and observation is bounded by $\pm 1$ mm/day. For the arid to semi-arid western Texas, both surface runoff and base flow are overpredicted by the model, with a positive bias less than $+0.1$ mm/day. However, given the dry climatology of this region, a bias of $+0.1$ mm/day bias is quite significant as it represents an error percentage of more than $+100\%$ with respect to the observed data (see Figures 8c and 8d). Overall, large mass-balance errors are evident in the arid west and the relatively dry Nueces River Basin in the southwest. This further explains the negative NSE values (Figure 6b) over the drier parts of Texas, where the mass-balance error dominates the level of model accuracy despite a generally reasonable capture of the temporal variability in streamflow (Figure 6a, CC ranging from 0.2 to 0.6 in western Texas). In the humid eastern Texas, the reasonable to good model performances are attributed to the relatively small percentage differences (within $\pm 50\%$). The only exception is seen over a northeast-southwest band in eastern Texas (Figure 6b, negative NSE values). Within this band, the low NSE values correspond well with the large base flow overprediction ($>+100\%$) problem, suggesting possible model improvements that are largely dependent on the base flow simulation in this area. Despite the general base flow overprediction issue, some slight underpredictions are also seen over the three wettest regions, including the San Jacinto River Basin, the upper Trinity River Basin, and the Guadalupe and San Antonio River Basins (generally within 0 to $-40\%$).

To understand the possible interplay between ET and two individual runoff components, we plot the bias in modeled surface runoff (Figure 9a) and base flow (Figure 9b) against the bias in ET averaged over 2008 to 2011. The term “bias” is used here under the assumption that the USGS runoff and the MODIS ET data provide the best available observational estimates, which may not be true due to data uncertainties. However, the rationale behind conducting this analysis is due to the emerging interests in using earth observations to calibrate hydrologic models at large spatial scales, where RS-ET products are considered as the best-available “observation” to improve streamflow simulations (Cai et al., 2014a; Kunnath-Poovakk et al., 2016). It is thus important to analyze how these biases in the mutually dependent hydrologic processes can be related with the theoretically expected model performances, before remotely sensed data should be introduced in model calibration.

In Figure 9, the ET bias and runoff bias are calculated as the drainage basin averages to represent physically-consistent quantities. Each dot represents
181 out of the 271 drainage areas of the USGS gauges, whose area is larger than 625 km² (5 × 5 LSM grid cells) to filter out some random noises that are present in small watersheds. The color scheme represents the AI at the 181 gauge locations, ranging from 0.2 to 0.6. Ideally, errors in ET and runoff (i.e., two outgoing fluxes from the land surface) are expected to show a complementary negative relationship to close the water balance. Due to data uncertainties and the limited number of years used, it is not surprising that such a perfect relationship is less obvious in Figure 9, similar to what is being documented in a recent study evaluating multiple LSMs (Zhang et al., 2016). However, Figure 9 still displays an interesting pattern where the base flow bias, compared to surface runoff bias, is more distinguishably influenced by the bias in ET (Figure 9b, a significant negative correlation). This is reasonable because base flow is more directly influenced by soil moisture and groundwater storages in which the ET partitioning plays a significant role. Accordingly, model calibration against MODIS ET may improve base flow simulation, as also implied by Kunnath-Poovakka et al. (2016). In comparison, the bias in surface runoff seems to be less sensitive to the bias in ET (Figure 9a). This is because generation of surface runoff is also closely related to rainfall intensity and land surface heterogeneity in addition to being influenced by soil wetness conditions. Although arguably ET can still affect surface runoff by controlling the antecedent soil moisture profile at the onset of a rainfall event, this cannot be ascertained from the current study due to the highly complex nonlinearity in the rainfall-runoff processes and data uncertainties. Future studies need to be conducted with various ET products to further understand this problem.

Figure 9 also displays some unexpected patterns where both the runoff and ET fluxes can be over-predicted (or underpredicted) in the same direction. Such patterns can be ascribed to data uncertainties, or it may also reflect model biases in the soil moisture and groundwater storages. Theoretically, when model-predicted storages have large positive (negative) biases, it is likely that both of the outgoing fluxes (i.e., runoff and ET) can be underpredicted (overpredicted) to close water balance in long-term averages. Figure 9 shows that more than half of the dry region gauges (with overpredicted ET, surface runoff, and base flow) and a few wet region gauges (with underpredicted ET and surface runoff) are subject to this problem. Therefore, at these locations, one might not expect improved model performances by simply calibrating the model using ET observations. Instead, more hydrologic observations (both in situ and RS-based) on soil moisture and water table depth (WTD) are needed to further improve the streamflow prediction. In our current model configuration, the Noah-MP runoff parameterization scheme is tightly coupled to a simple 1D groundwater model with an unconfined aquifer (Niu et al., 2007), in which runoff is parameterized as an exponential decay function of the model-simulated WTD. Under this parameterization scheme, Cai et al. (2014a) reported a reasonably simulated spatial pattern of WTD in their study domain, yet it was noted that the magnitude of WTD was simulated too shallow. In an attempted analysis to compare modeled WTD with the Texas Water Development Board well measurements (not shown), we found that the simulated WTD is indeed too shallow especially in the arid western Texas, which is likely related to the underrepresented subsurface properties. This may help explain the overpredicted ET, surface runoff, and base flow in most of dry region gauges according to the Noah-MP runoff parameterization. However, this study is limited in proving the hypothesis, which warrants further investigation in this direction. Besides, this study is based on a four-year model simulation and one source of the RS ET product. More general and robust conclusions may benefit from multiple long-term simulations and quantitative analyses with different data products in future studies.
CONCLUSIONS

Uncertainty in streamflow prediction from a large-scale hydrologic forecasting system mainly originates from meteorological forcing, as well as the representations of land surface and river routing processes. This study is an attempt to evaluate the spatiotemporal performance of a offline WRF-Hydro-RAPID modeling experiment by focusing on its ET and streamflow simulations over Texas. First, by comparing model simulated ET and MODIS ET, we found that, spatially, the modeled ET performance is aridity-dependent; temporally, the ET fluctuations are better captured in a normal-to-wet year than in an extreme-drought year. Second, by comparing model-simulated streamflow against 271 gauge observations, a more complete picture of the model performance is provided to complement previous studies that mainly focused on major rivers and basin outlets. In Texas, streamflow is better predicted in the wet region than in the dry region, with the highest daily NSE at ~0.7 based on uncalibrated results. In comparison, presence of the large positive runoff bias over the dry region indicates a persistent problem in simulating the mass-balance components by the model. Third, by applying a recursive digital filter to decompose runoff components, the results provide a large-scale evaluation of the surface and base flow simulations by Noah-MP, the LSM in the WRF-Hydro framework. We found that surface runoff is underpredicted while base flow is over-predicted in the wet region. In the dry region, both surface runoff and base flow are largely over-predicted, which results in the poor model performance there. Finally, by linking the errors (the differences calculated as the model estimates minus the reference estimates) between ET and the two runoff components, we found that base flow errors have a stronger correlation with ET errors than in the case for surface runoff. This has implications for large-scale model calibration studies that use RS-ET data to improve streamflow predictions. At most of the dry region gauges where runoff and ET errors do not show complementary relationships, calibrating ET may not yield improved streamflow predictions. In these situations, possible solutions to model improvements focusing on better physical parameterizations of soil moisture and/or groundwater might render more useful solutions through future endeavors.

Although this study was conducted under the scope of the 2015 NFIE, our goal spanned beyond flood prediction with a view to provide a holistic assessment on the model’s overall hydrologic performance. Accordingly, the findings would complement the ongoing improvements of the NWM, a continental-scale high resolution flood forecasting framework operationally running for the U.S. (http://water.noaa.gov/about/nwm). Given that our uncalibrated modeling results under a much simpler model configuration than that adopted by the NWM still produce promising results for a range of gauge locations, and considering there are continuous efforts by the NWM development team that focuses on reducing the uncertainty in streamflow prediction from all error sources (e.g., from atmospheric forcing to human influences), it is expected that the operational hydrologic prediction skills of the nation will be enhanced in the future. Our study, by assessing the large-scale fine-resolution ET and streamflow simulations over Texas, is one important initial step towards accurate hydrologic forecasts.

ACKNOWLEDGMENTS

This work was funded by the National Natural Science Foundation of China grant 41375088, the NSF Coupled Natural and Human Systems Program award 1518541, the Cynthia and George Mitchell Family Foundation, the Texas Water Research Network, and Microsoft Research. The authors are grateful to the CUAHSI funding support for the Summer Institute (June 1 to July 18, 2015) held at the National Water Center, Alabama, U.S. We thank David Gochis (National Center for Atmospheric Research), Jim Nelson (Brigham Young University), and Fernando Salas (NOAA National Water Center) for providing project advice as NFIE advisors and course coordinators. David Arctur (University of Texas at Austin) is thanked for English editing. We would also like to thank the Editor-in-Chief Jim Wigington, Associate Editor David Tarboton (Utah State University), and three anonymous reviewers for their constructive comments.

LITERATURE CITED
