Comprehensive Evaluation of the Variable Infiltration Capacity (VIC) Model in the North American Land Data Assimilation System^①

YOULONG XIA,^{a,b} DAVID M. MOCKO,^{c,d} SHUGONG WANG,^{c,d} MING PAN,^e SUJAY V. KUMAR,^c CHRISTA D. PETERS-LIDARD,^f HELIN WEI,^{a,b} DAGANG WANG,^g AND MICHAEL B. EK^{a,h}

^a NOAA/National Centers for Environmental Prediction/Environmental Modeling Center, College Park, Maryland

^b I.M. Systems Group, and NOAA/NCEP/EMC, College Park, Maryland

^c Hydrological Sciences Laboratory, NASA Goddard Space Flight Center, Greenbelt, Maryland

^d Science Applications International Corporation, and NASA GSFC, Greenbelt, Maryland

^e Department of Civil and Environmental Engineering, Princeton University, Princeton, New Jersey

^f Earth Science Division, NASA Goddard Space Flight Center, Greenbelt, Maryland

^g Department of Water Resources and Environment, Sun Yat-Sen University, Guangzhou, China

(Manuscript received 6 July 2018, in final form 3 October 2018)

ABSTRACT

Since the second phase of the North American Land Data Assimilation System (NLDAS-2) was operationally implemented at NOAA/NCEP as part of the production suite in August 2014, developing the next phase of NLDAS has been a key focus of the NCEP and NASA NLDAS teams. The Variable Infiltration Capacity (VIC) model is one of the four land surface models of the NLDAS system. The current operational NLDAS-2 uses version 4.0.3 (VIC403), the research NLDAS-2 used version 4.0.5 (VIC405), and the NASA Land Information System (LIS)-based NLDAS uses version 4.1.2.1 (VIC412). The purpose of this study is to evaluate VIC403 and VIC412 and check if the latter version has better performance for the next phase of NLDAS. Toward this, a comprehensive evaluation was conducted, targeting multiple variables and using multiple metrics to assess the performance of different model versions. The evaluation results show large and significant improvements in VIC412 over the southeastern United States when compared with VIC403 and VIC405. In other regions, there are very limited improvements or even deterioration to some degree. This is partially due to 1) the sparseness of USGS streamflow observations for model parameter calibration and 2) a deterioration of VIC model performance in the Great Plains (GP) region after a model upgrade to a newer version. Overall, the model upgrade enhances model performance and skill scores for most parts of the continental United States; exceptions include the GP and western mountainous regions, as well as the daily soil moisture simulation skill, suggesting that VIC model development is on the right path. Further efforts are needed for scientific understanding of land surface physical processes in the GP, and a recalibration of VIC412 using reasonable reference datasets is recommended.

1. Introduction

The multi-institution North American Land Data Assimilation System (NLDAS) was operationally implemented at NOAA/NCEP in August 2014. Since then, its

upgrade and development has continued to be a critical task for the NCEP and NASA NLDAS teams. Recently, the two teams have used the NASA Land Information System (LIS; Kumar et al. 2006) software framework to run several potential model candidates to prepare for the next phase NLDAS implementation. These models include version 3.6 of the Noah land model (Ek et al. 2003), version 3.6 of Noah-MP (Niu et al. 2011), the Fortuna-2.5 version of the NASA Catchment model (Koster et al. 2000), and version 4.1.2.1 of the Variable Infiltration Capacity (VIC) model (Liang et al. 1994; Newman et al. 2017). The VIC model is a physically based hydrological model and has been continuously developed and upgraded. Therefore, many different model versions exist. These upgrades include improvements to the model

^① Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/JHM-D-18-0139.s1>.

^h Current affiliation: Joint Numerical Testbed, Research Applications Laboratory, National Center for Atmospheric Research, Boulder, Colorado.

Corresponding author: Youlong Xia, youlong.xia@noaa.gov

physics as well as model parameters so that various water and energy fluxes/states can be better simulated. The terrestrial water storage and its individual components of some of these updated LSMs were evaluated by [Xia et al. \(2017\)](#). However, many other variables such as energy budget components, water budget components, and other state variables have not yet been evaluated against observations or other types of reference data.

For decades, the NCEP NLDAS team and its partners have employed a variety of observations and reference datasets for various evaluation studies (e.g., [Robock et al. 2003](#); [Mitchell et al. 2004](#); [Lohmann et al. 2004](#); [Xia et al. 2012a](#); [Kumar et al. 2014, 2017](#)). These include site measurements such as soil temperature, soil moisture, and USGS streamflow. The gridded reference datasets include NASA Global Energy and Water Exchanges project (GEWEX) Surface Radiation Budget (SRB) monthly radiation fluxes ([Zhang et al. 2013, 2015](#)), multitree ensemble FLUXNET-based (MTE) monthly sensible and latent heat fluxes generated by [Jung et al. \(2009\)](#), USGS monthly Hydrologic Unit Code 8 (HUC8) total runoff ([Velpuri et al. 2013](#)), MTE monthly evapotranspiration (ET; [Jung et al. 2009](#)), and Global Land Evaporation Amsterdam Model (GLEAM) monthly ET ([Miralles et al. 2011](#); [Martens et al. 2017](#)). Monthly soil temperature was measured from 137 cooperative stations ([Xia et al. 2013](#)); monthly soil moisture was measured from 17 sites in Illinois ([Robock et al. 2000](#)), 97 sites in Oklahoma ([Scott et al. 2013](#)), and 59 sites in West Texas ([Schroeder et al. 2005](#)). The gridded products also include monthly snow water equivalent from the Snow Data Assimilation System (SNODAS; [Clow et al. 2012](#)) and Gravity Recovery and Climate Experiment (GRACE) monthly terrestrial water storage anomaly ([Landerer and Swenson 2012](#)). These data have been used in NLDAS-2 system development and evaluation. However, they are usually used separately to evaluate one or several individual variables, instead of all water budget components, energy budget components, and state variables. For more details, see the NLDAS product validation website (<https://ldas.gsfc.nasa.gov/nldas/NLDAS2valid.php>) and publication website (<https://ldas.gsfc.nasa.gov/nldas/NLDASpublications.php>). Compared with other NLDAS-2 models, the VIC model shows better performance for streamflow, total runoff, and SWE simulation, although VIC has a fair performance for ET simulation. For soil moisture simulation, its performance is comparable to other NLDAS-2 models for all soil layers except for the top 10-cm soil layer where the VIC model has lower seasonal variability ([Xia et al. 2014](#)). Overall, the VIC hydrologic model is a valuable member of the NLDAS model ensemble. Internally, the NLDAS team has over

the past few years developed the NLDAS Science Testbed to evaluate the new LSMs—including various versions, model options/configurations, and parameters—against numerous available budget components and variables; however, much of the previously published work did not include a comprehensive evaluation including multiple variables and metrics. In addition, most of our previous evaluation work did not include a statistical significance test (e.g., [Zou 2007](#); [Silver et al. 2015](#)), so we cannot demonstrate whether either the model version upgrade or model parameter upgrade leads to significant improvement/deterioration in the model performance. Recent VIC model development in the NLDAS system provides an opportunity to allow us to establish a comprehensive evaluation framework to support development of the next phase of NLDAS ([Ek et al. 2017](#)) and to guide NLDAS-like system developers. It should be noted that different time periods will be used in order to match these in situ and satellite observations, as well as other reference datasets.

The evaluation includes multiple metrics and tests, whether or not these metrics are statistically different between model versions, including bias/relative bias (Student's *t* test), root-mean-square error (RMSE), Taylor skill score ([Taylor 2001](#)), and anomaly correlation ([Zou 2007](#)), for test references used in past decades for NLDAS development. It also analyzes the roles of model version upgrades and calibrated model parameters in VIC model development. The paper is organized as follows. After the model and data description in [section 2](#), the evaluation strategy is described in [section 3](#). The results, discussion, and future work are given in [section 4](#). The summary and conclusion are presented in [section 5](#).

2. Model and data

a. Model

The VIC model ([Liang et al. 1994](#)) is used in both the operational (version 4.0.3) and research (version 4.0.5) NLDAS-2 systems ([Xia et al. 2016a,b](#)). The VIC4.0.3 (hereafter called VIC403) was found to partition too much precipitation into runoff, particularly in the southeastern United States, when compared with the USGS streamflow observations from 1145 USGS small to medium ($<10\,000\text{ km}^2$) basins ([Lohmann et al. 2004](#)). That is due to suboptimal soil and hydrologic parameters. To reduce the large runoff biases in VIC403, [Troy et al. \(2008\)](#) conducted a multicriteria calibration for VIC403 by using daily Nash–Sutcliffe efficiency (NSE; [Nash and Sutcliffe 1970](#)) in the eastern United States (east of 100°W) and mean absolute error in the western United States (west of 100°W), between VIC and the

observed streamflow during October 1997 and September 2001 from 1131 small basins. Troy et al. (2008) then used the observations from October 2001 to September 2005 to validate the calibrated VIC403. The results showed that the runoff biases were significantly reduced. The VIC403 simulation used in this study is from the operational configuration, which used the uncalibrated parameters. The calibrated soil and hydrologic parameters (Troy et al. 2008) were then used in VIC405 and incorporated into the NCEP research NLDAS-2 system. Recently, these calibrated soil and hydrologic parameters were also used in VIC4.1.2.1 (hereafter called VIC412) to prepare the next phase NLDAS system upgrade. It should be noted that current NLDAS framework is used to run VIC403 and VIC405, and NASA's LIS (Kumar et al. 2006) software framework is used to run VIC412.

Besides the differences in soil and hydrologic parameters, the versions of the VIC model contain many other changes and upgrades. A detailed model development overview can be found on the University of Washington VIC website (<http://www.hydro.washington.edu/Lettenmaier/Models/VIC/Overview/ModelOverview.shtml>) and a technical note from Gao et al. (2009). Here, only a brief description of the significant changes is given. The VIC model has a range of configuration options, which may lead the model to have different parameterization schemes for the same land surface process. For example, VIC supports two baseflow schemes: one is based on the ARNO model (Liang et al. 1994) and the other is developed by Nijssen et al. (2001). Therefore, it is important to include model configurations when comparing the simulation results of different VIC versions. In this study, all of the VIC versions (VIC403, VIC405, VIC412) share the common configurations, including multiple land cover tiles per grid cell, single soil column per grid cell consisting of three layers, water balance and energy balance, modeling both surface snowpack and snow intercepted by canopy, and multiple snow elevation bands within a grid cell. In VIC403 and VIC405, distributed precipitation is turned on while it is turned off in VIC412 to avoid a water imbalance issue (Xia et al. 2012a). VIC403 and VIC405 share exactly the same parameterizations for all physical processes, and the only difference is that VIC405 has fixed a few "bugs" from VIC403, including to add more thorough checks on parameters and initial conditions to support user debugging and to reset out-of-range forcing values (e.g., negative shortwave). Although bug fixes may have some impacts on model simulations, we assumed that these impacts are negligible. A simple diagram

summarizes the three configurations of VIC used in this study (Fig. 1). When the VIC model was upgraded to version 4.1.2, many new features were added. This includes canopy temperature and snow air temperature calculations, parameterization of spatially varied snow cover and frozen soil and their impact on the snowmelt calculation, and calculation of blowing snow sublimation. The major update is related to snowpack simulation and land skin temperature calculation. The VIC model consists of two modes: a water mode for a water balance simulation and an energy mode for both water and energy balance simulations.

The VIC model is used in this study to simulate both land surface water and energy balances. The surface water and energy balance equations can be represented as

$$\frac{d(\text{SMC} + \text{CWS} + \text{SWE})}{dt} = P - Q - \text{ET}, \quad (1a)$$

$$\text{SWNET} + \text{LWNET} = \text{SH} + \text{LH} + \text{GS}, \quad (1b)$$

where SMC is total column soil moisture content, CWS is canopy water storage, SWE is snow water equivalent, P is surface precipitation, Q is total runoff, and ET is evapotranspiration. SWNET is net shortwave radiation, LWNET is net longwave radiation, SH is sensible heat flux, LH is latent heat flux, and GS is ground heat flux. The model uses the forcing data as well as soil and hydrologic parameters to simulate spatial and temporal variation of these two balances.

b. Model setup

The spinup procedure for VIC403 and VIC405 strictly follows the NLDAS-2 system requirement (Xia et al. 2012a). The 15-yr (1979–94) spinup time is used to run the VIC model to achieve the equilibrium state. The criterion for the VIC model to reach equilibrium state is that monthly soil moisture means at deep soil layer are within 0.01% of each other from year to year for 99% of all grid cells (Cosgrove et al. 2003). For VIC412, a 70-yr spinup time is used, with the model run twice for a 35-yr period from January 1979 to December 2013 to achieve the equilibrium state. Other new LSMs for the next phase of NLDAS include a groundwater module and need a longer spinup time, to keep consistency between the new models. In spite of the different spinup times for the versions of VIC in this study, in an operational viewpoint, the impact of initial states on simulations should be negligible, especially considering that all simulations start from January 1979 with their respective equilibrium states, and period of the reference data for the evaluation is mostly from the 2000s, or at least

FIG. 1. A schematic diagram of the versions of the VIC model as configured for the simulations in this study. Details about some of the model physics changes can be obtained from the VIC model website (see <http://www.hydro.washington.edu/Lettenmaier/Models/VIC/Development/ArchivedVersions.shtml#RecentDev>).

starting from the mid-1980s. All three versions of the VIC models use the same NLDAS-2 hourly surface meteorological forcing data including gauge-based precipitation, downward shortwave radiation, downward longwave radiation, 2-m air temperature and specific humidity, 10-m wind speed, and surface pressure (Xia et al. 2012a,b). The soil texture, vegetation classification and fraction, and other soil and vegetation related parameters [other than the calibrated parameters from Troy et al. (2008)] strictly follow the current operational NLDAS-2 system setup, as this new test will be potentially used to upgrade to the next phase NLDAS system. For soil and hydrologic parameters, VIC405 and VIC412 use the calibrated parameters (Troy et al. 2008), which are different from those used in VIC403.

c. Data

The reference datasets used in this study are summarized in Table 1. The gridded products consist of energy fluxes, water fluxes, and state variables (e.g., snow water equivalent, terrestrial water storage). The energy fluxes include monthly net shortwave radiation, monthly net longwave radiation, monthly sensible heat flux, and monthly latent heat flux. The water fluxes include NLDAS-2 hourly precipitation, monthly USGS HUC8 runoff, MTE monthly ET, and GLEAM daily ET. Two reference ET products are selected as they have a long-term record when compared with the other satellite-based ET products. Furthermore, recent evaluations conducted at both flux tower site and global scales show that GLEAM-based ET is superior to other satellite-based ET products (Michel et al. 2016; Miralles et al. 2016). It should be noted that MTE ET and

GLEAM are generated from observations and model outputs and thus inevitably include errors. The MTE ET product provides the uncertainty estimate (Jung et al. 2009, 2011) to help assess various errors (e.g., forcing errors, model structure errors). Generally, MTE ET has more accurate mean annual ET and spatial pattern but underestimates interannual variability (e.g., amplitude) when compared with the GLEAM product. GLEAM has larger errors over the central region when compared with other regions in the continental United States (CONUS) (Miralles et al. 2011). Although such remote sensing ET data also contain errors, as an independent data source, they have been widely used to evaluate land surface model ET simulations (Jiménez et al. 2011; Mueller et al. 2011; Xia et al. 2016a; Kumar et al. 2018a). The state variables include monthly terrestrial water storage anomaly and daily snow water equivalent. In addition, the state variables also include in situ point observations such as 1) monthly soil temperature from 137 NOAA cooperative stations; 2) monthly soil moisture in Illinois from their soil moisture network and daily soil moisture from the Oklahoma and West Texas Mesonets (for dense measurement networks); and 3) daily soil moisture from the USDA Agricultural Research Service (ARS), Soil Climate Analysis Network (SCAN), and U.S. Climate Reference Network (USCRN) over the CONUS (for a more widely distributed network over a range of climatic conditions). The details can be found in Table 1 and Fig. 2. For evaluation purposes, the hourly and daily data are aggregated to monthly time scales. As these products have different spatial resolutions, all products are regridded to the NLDAS-2 grid at 0.125° spatial resolution as done by

TABLE 1. Description of the reference datasets used in this study.

Variable name	Data period	Spatial resolution/No. of stations	Temporal resolution	Data source and website	Reference
Net shortwave radiation (SWNET)	1984–2007	Gridded products 1°	Monthly	NASA GEWEX SRB website: https://gewex-srb.larc.nasa.gov/	Zhang et al. (2013)
Net longwave radiation (LWNET)	1984–2007	1°	Monthly	Same as above	Zhang et al. (2015)
Sensible heat flux (SH)	1984–2007	0.5°	Monthly	Max-Planck Institute for Biogeochemistry: https://www.bgc-jena.mpg.de/bgi/index.php/Services/Overview	Jung et al. (2009)
Latent heat flux (LW)	1984–2007	0.5°	Monthly	Same as above	Xia et al. (2012a)
NLDAS-2 precipitation (<i>P</i>)	1982–2011	0.125°	Hourly	NCEP NLDAS website: http://www.emc.ncep.noaa.gov/mmb/nldas/	Velpuri et al. (2013)
USGS runoff (<i>Q</i>)	1982–2011	~62 km	Monthly	USGS website: https://waterwatch.usgs.gov/index.php?id=romap3&sid=w_download	Jung et al. (2009)
MTE Evapotranspiration (ET)	1982–2011	0.5°	Monthly	Max-Planck Institute for Biogeochemistry	Miralles et al. (2011)
GLEAM Evapotranspiration (ET)	1982–2011	0.25°	Daily	http://www.gleam.eu/#datasets	Clow et al. (2012)
Snow water equivalent (SWE)	2004–13	0.01°	Daily	NOHRSC SNODAS website: https://www.nohrsc.noaa.gov/archived_data/	Landerer and Swenson (2012)
Terrestrial water storage anomaly (TWSA)	2003–12	1°	Monthly	NASA GRACE website: https://grace.jpl.nasa.gov/data/get-data/	Xia et al. (2013)
CONUS soil temperature (5, 25, 70 cm)	1979–2001	In situ observations 137	Monthly	ftp://ldas.ncep.noaa.gov/Youlong.Xia/SoilT/archive	Robock et al. (2000)
Soil moisture in IL (5, 25, 70, 150, 100, 200 cm)	1985–2004	17	Monthly	NASMD website: http://nationalsoilmoisture.com/About.html	Scott et al. (2013)
Soil moisture in OK (5, 25, 70 cm)	2000–10	97	Daily	Same as above	Schroeder et al. (2005)
Soil moisture in West Texas (5, 25, 70 cm)	2000–10	59	Daily	Same as above	Jackson et al. (2010)
Soil moisture in ARS (5 cm)	2001–16	7	Daily	Same as above	Schaefer et al. (2007)
Soil moisture in SCAN (5, 25, 70 cm)	2000–17	119	Daily	Same as above	Bell et al. (2013)
Soil moisture in USCRN (5, 25, 70 cm)	2010–17	86	Daily	Same as above

FIG. 2. (a) Locations of 137 soil temperature measurement stations (filled circles), 17 soil moisture measurement stations in Illinois, 97 soil moisture measurement stations in Oklahoma, and 59 soil moisture measurement stations in West Texas (open circles). (b) Locations of 7 USDA ARS soil moisture measurement sites, 119 USDA SCAN measurement sites, and 86 NOAA USCRN sites.

many previous evaluation works (Velpuri et al. 2013; Cai et al. 2014; Xia et al. 2016a, 2017). In this study, a mass conservative bilinear interpolation method is used for ET products; a bilinear interpolation method is used for downward shortwave and longwave radiation; and the nearest-neighbor method is used for soil moisture, soil temperature, and snow water equivalent. Most of the products in Table 1 have been used in previous validation work for NLDAS-2 development and application (Xia et al. 2013, 2014, 2015a, 2016a,b, 2017).

Because of different data periods in the various reference datasets, we generally use a common period for energy flux evaluation (January 1984–December 2007) and for water flux evaluation (January 1982–December 2011). However, it is difficult to find a common period for all the state variables evaluated. Therefore, we decided to use different periods for different state variables. The Snow Data Assimilation System (SNODAS) SWE from January 2004 to December 2013 and the GRACE terrestrial water storage change anomaly

FIG. 3. Spatial distribution of 30-yr (1982–2011) averaged annual reference values calculated from (a) USGS Q , (b) water-budget-derived ET, (c) MTE FLUXNET ET, and (d) GLEAM ET datasets, (e)–(h) Rbias between VIC403 and the references, and (i)–(l) Rbias between VIC412 and the references. Areas with the insignificant Rbias at the 95% confidence level (Student's t test) are masked out.

(TWSCA) from January 2003 to December 2012 are used. For state variable evaluation purposes, 23-yr (1979–2001) monthly soil temperature in United States, 20-yr (1985–2004) monthly soil moisture in Illinois, 11-yr (2000–10) monthly soil moisture in Oklahoma, 11-yr (2000–10) monthly soil moisture in West Texas, 16-yr (2001–16) daily soil moisture in the USDA ARS network, 15-yr (2000–14) daily soil moisture in the SCAN network, and 8-yr (2010–17) daily soil moisture in USCRN network are also used. The monthly modeled water fluxes, energy fluxes, and state variables from the VIC403, VIC405, and VIC412 model simulations are obtained from the NLDAS project (Xia et al. 2016a,b, 2017) for the corresponding periods as mentioned above.

3. Evaluation strategy

In this study, we use multiple metrics and multiple reference datasets to evaluate VIC403, VIC405, and VIC412. The primary purpose is to evaluate the impact of updating to a new model version or implementing a

parameter calibration procedure in the NLDAS configuration. Therefore, the evaluation mainly focuses on VIC403 and VIC412. The VIC405 simulation is used only as a sensitivity test in the discussion section.

All variables in Eqs. (1a) and (1b) except for CWS and GS (due to limited availability of reference datasets) are evaluated against either in situ observations or other reference datasets. In addition, soil temperature and terrestrial water storage change [$TWSC = d(SMC + CWS + SWE)/dt$] are also evaluated against in situ observations and GRACE satellite retrievals, respectively. These variables generally cover almost all energy and water cycle features in the land surface model. Some of them are also used for drought monitoring: for example, P is used to monitor meteorological drought, Q is used to monitor hydrological drought, SMC and ET are used to monitor agricultural drought, SWE is used for winter and spring drought monitoring, and the TWSC anomaly is used for overall drought monitoring. Such a multivariate analysis can provide a complete evaluation picture for the NLDAS system upgrade procedure.

FIG. 4. Spatial distribution of 24-yr (1984–2007) averaged annual reference values calculated from NASA GEWEX/SRB (a) SWNET and (b) LWNET and MTE FLUXNET (c) SH and (d) LH datasets, (e)–(h) Rbias between VIC403 and the references, and (i)–(l) Rbias between VIC412 and the references. The insignificant Rbias at the 95% confidence level (Student's t test) is masked out.

The metrics used in this study are bias and relative bias (Rbias), RMSE, anomaly correlation (AC), and Taylor skill score S . The bias and Rbias are used to evaluate model systematic error, RMSE is used to evaluate the overall model error, AC is used to evaluate model skill and temporal variability, and S is used to evaluate the overall skill. These metrics have been widely used by the model evaluation community (for more details, see supplemental material). In this study, we used difference metrics (dRMSE, dS, dAC) to quantify the improvement (i.e., negative dRMSE and positive dS and dAC) or deterioration (positive dRMSE and negative dS and dAC) of model skills in different VIC versions. Significance tests were also performed on these differences. For the difference in annual mean values between two VIC versions, a Student's t test is done using the FORTRAN code from Press et al. (1992). For AC differences between two VIC versions, an approach developed by Zou (2007) and implemented by Silver et al. (2015) is used. The Zou approach computes the confidence intervals for dependent datasets to decide if two

correlations are significantly different. We use a 95% confidence level for the mean difference and dAC analysis. As indicated by Taylor (2001), a decrease in RMSE may not necessarily be considered an improvement in skill when the correlation coefficient is relatively low. Therefore, the use of a multivariate evaluation may avoid some misleading results. However, due to the use of different data sources as the reference, water and energy budget imbalance errors exist, which may result in conflicting conclusions in some cases. This needs to be cautiously interpreted so that a more robust conclusion can be drawn.

4. Evaluation results

a. Mean annual data

1) WATER BUDGET COMPONENTS

The 30-yr (1982–2011) mean annual Q and ET spatial distributions over the continental United States have a common feature: large Q and ET exist in the eastern

United States and along the West Coast while small ET and Q exist in the interior United States (Figs. 3a–d). We also calculate the mean annual Rbias for Q and ET when several different references are used. For the Q evaluation, mean annual regridded USGS HUC8 runoff is used as the reference. For the ET evaluation, we use three reference datasets, as these references may have large uncertainties. The first reference ET is calculated from the water budget method ($ET = P_{NLDAS} - Q_{USGS}$), the second one is the FLUXNET MTE product, and the third one is the GLEAM product. The results show that Rbias of Q is significantly reduced in the eastern United States, the Southeast in particular (Figs. 3e,i), when comparing VIC412 with VIC403, which is consistent with the result from Troy et al. (2008), although they used short-term streamflow data to calibrate the VIC403. The large Rbias is also significantly reduced in the ET simulation for the eastern United States as precipitation is partitioned between ET and Q when the same precipitation is used to drive VIC412 and VIC403 (Figs. 3f–h and 3j–l). However, in the western United States, there are very limited improvements for both Q and ET in North Dakota and northern Nebraska. One reason is that when VIC403 is calibrated for the western United States, its bias is moderately reduced for only a few basins. In addition, having fewer observations in the western United States limits the VIC model calibration procedure and this leads to less improvement in that region.

2) ENERGY BUDGET COMPONENTS

The map of mean net shortwave and longwave shows higher values in the southwest and lower values in the northeast (Figs. 4a,b). There are high sensible heat flux values in the west and low values in the east, while the latent heat flux is high in the east and low in the west (Figs. 4c,d). As the same downward shortwave (DSWR) and longwave radiation (DLWR) are used as forcing data for both VIC403 and VIC412, the major differences come from the upward shortwave (USWR) and longwave radiation (ULWR). The USWR can be calculated as

$$\begin{aligned} \text{USWR} = & (1.0 - F_{\text{snow}})\alpha_{\text{surface}}\text{DSWR} \\ & + F_{\text{snow}}\alpha_{\text{snow}}\text{DSWR}, \end{aligned} \quad (2)$$

where F_{snow} is snow cover fraction, and α_{snow} and α_{surface} are snow and snow-free surface albedo, respectively. As the snow-free surface albedo did not change between the versions, the difference comes from the snow cover areas simulated from VIC412 and VIC403. A significant net shortwave radiation bias reduction appears in the western high mountain region (Figs. 4e,i) likely due to the upgrade of snowpack related processes in VIC412 (Fig. 1). The ULWR can be calculated as

$$\text{ULWR} = \varepsilon\sigma T_{\text{skin}}^4, \quad (3)$$

where ε is the land surface emissivity which is set to 1, σ is the Stefan–Boltzmann constant, and T_{skin} is the land surface skin temperature. The major difference in net longwave radiation comes from the land surface skin temperature calculated from snow and snow-free surface temperature which is weighted by snow cover fraction. The analysis (Figs. 4f,j) shows that negative Rbias is changed into positive Rbias in the regions with more snow cover such as the Northeast, Great Lakes, and western mountain region due to a VIC model upgrade, which includes snowpack-related processes and calculations of snow and canopy temperature.

A significant Rbias decrease is found for SH (Figs. 4g,k) and LH (Figs. 4h,l) in the southeastern United States. In the northeastern, Great Lakes, and northwestern regions, Rbias is somewhat increased for SH and LH due to a significant change in net longwave radiation. This suggests that the land surface skin temperature in these regions may be too cold in VIC412. A small ULWR change leads to a large net longwave radiation as the downward longwave radiation is the same.

3) SWE, SOIL TEMPERATURE, AND SOIL MOISTURE

The SWE Rbias analysis shows that both VIC412 and VIC403 underestimate the SWE if compared with SNODAS for most regions in United States (Figs. 5a,b), suggesting the need for further improvement. This large negative Rbias can be attributed to three factors: 1) large SNODAS SWE (Clow et al. 2012); 2) NLDAS-2 forcing errors such as light precipitation, partitioning total precipitation into rainfall and snowfall, and warm air temperatures; and 3) inefficiency in VIC model snowpack-related parameterizations and model parameters. The detailed discussion can be found in section 5d. However, when comparing VIC412 and VIC403 (Fig. 5c), the upgrade to snowpack-related processes does lead to a significant SWE increase in the Northeast, Great Lakes, and western mountains. This indicates that the model upgrade is moving in the right direction. The RMSE is generally reduced ($d\text{RMSE} < 0$) and dS and dAC are increased ($dS > 0$, $dAC > 0$) in most part of high latitudes and mountainous regions, suggesting simulated SWE improvement when comparing VIC412 with VIC403 (Figs. 5d–f). However, there are still some positive $d\text{RMSE}$ and negative dS values in many regions, indicating SWE skill degradation in VIC412, which suggests the need for further investigation.

Soil temperature bias in VIC412 is reduced for all three soil layers when compared with VIC403, although

FIG. 5. Spatial distribution of SWE bias (mm month⁻¹) between 10-yr (2004–13) averaged (a) VIC403 and SNODAS, (b) VIC412 and SNODAS, and (c) VIC412 and VIC403; (d) dRMSE (VIC412-VIC403; mm month⁻¹), (e) dS (VIC412-VIC403; unitless), and (f) dAC (VIC412-VIC403; unitless). The insignificant Rbias (Student's *t* test) and dAC (Zou test) at the 95% confidence level (Student's *t* test) are masked out.

TABLE 2. Metrics for soil temperature averaged over 137 stations (ST1/0–10 cm, ST2/10–40 cm, ST3/40–100 cm) and moisture in Illinois, Oklahoma, and West Texas (SM1/0–10 cm, SM2/10–40 cm, SM3/40–100 cm, top 1 m, top 2 m). The station locations are plotted in Fig. 2. Bold numbers indicate significance to 95% (italic fonts indicate improvement, and normal fonts indicate degradation). The (s.) indicates significant and (n.s.) indicates not significant.

Variable	RMSE		Bias		AC		S	
	VIC 403	VIC412	VIC 403	VIC412	VIC 403	VIC412	VIC403	VIC412
CONUS averaged soil temperature statistics (K)								
ST1	0.64	0.76	0.25	0.13 (n.s.)	0.89	0.89 (n.s.)	0.998	0.995
ST2	0.99	0.85	0.69	0.38 (n.s.)	0.82	0.81 (n.s.)	0.994	0.996
ST3	2.08	1.35	1.66	0.90 (n.s.)	0.72	0.71 (n.s.)	0.963	0.989
Illinois averaged soil moisture statistics								
SM1	0.079	0.085	−0.004	0.035(s.)	0.614	0.622(n.s.)	0.066	0.077
SM2	0.081	0.039	0.000	0.016(s.)	0.826	0.719(s.)	0.701	0.850
SM3	0.057	0.038	−0.049	−0.031(s.)	0.830	0.802(s.)	0.660	0.784
SM4	0.093	0.117	−0.078	−0.115(s.)	0.662	0.717(s.)	0.103	0.725
Top 1 m	0.034	0.025	−0.030	−0.010(s.)	0.854	0.788(s.)	0.956	0.918
Top 2 m	0.061	0.065	−0.054	−0.062(s.)	0.899	0.803(s.)	0.578	0.788
Oklahoma averaged volumetric soil moisture statistics								
SM1	0.046	0.081	0.039	0.078(s.)	0.381	0.343(s.)	0.429	0.364
SM2	0.041	0.046	0.033	0.038(n.s.)	0.310	0.328(n.s.)	0.541	0.757
SM3	0.035	0.041	0.030	0.038(n.s.)	0.596	0.624(n.s.)	0.735	0.795
Top 1 m	0.040	0.038	0.035	−0.034(n.s.)	0.503	0.397(n.s.)	0.651	0.747
West Texas averaged volumetric soil moisture statistics								
SM1	0.089	0.119	0.079	0.111(n.s.)	0.540	0.383(s.)	0.291	0.303
SM2	0.066	0.048	0.057	0.037(n.s.)	0.744	0.658(n.s.)	0.417	0.514
SM3	0.057	0.066	0.051	0.057(n.s.)	0.900	0.784(s.)	0.813	0.470
Top 1 m	0.051	0.060	0.047	0.053(n.s.)	0.884	0.765(s.)	0.900	0.548

this reduction is not significant (Table 2). For soil moisture analysis in three regions (Illinois, Oklahoma, and West Texas), the bias in VIC412 is somewhat increased and the AC is reduced for most cases. For 5, 25, and 150 cm, and the top 2-m soil moisture in Illinois and 5-cm soil in Oklahoma, VIC412 significantly increases the biases when compared with VIC403, although it significantly reduces biases in the 70 cm and top 1-m soil layers in Illinois. In general, the first two layers of soil moisture in VIC412 become wetter and the bottom soil moisture layer becomes drier than in VIC403 due to the calibrated soil and hydrologic parameters (Troy et al. 2008).

b. Mean monthly data

1) MEAN SEASONAL CYCLE

Figure 6 shows the mean seasonal cycle of ET (Fig. 6a), Q (Fig. 6b), SWNET (Fig. 6c), LWNET (Fig. 6d), SH (Fig. 6e), and LH (Fig. 6f) in the southeastern United States (67–90°W, 25–40°N) when simulations and reference data are compared. The southeastern United States is selected because it has a large Rbias reduction, as analyzed in section 4a. For the ET evaluation, we used the MTE and GLEAM ET as the reference to represent uncertainty; for the Q ,

SWNET, and LWNET evaluations, we use an error bar (one standard deviation) to represent the reference data uncertainty; for the SH, LH, and ET evaluations, we use the MTE uncertainty estimate to represent data uncertainty (yellow shaded area). The results show that in VIC412, ET (SH) increases (decreases) and Q (LH) decreases (increases) when compared with VIC403. There are few changes for SWNET and LWNET when VIC412 and VIC403 are compared. Large improvements appear in the warm season (April–September) and small improvements appear in the cold season (October–March). When mean seasonal cycles are calculated over the United States domain (Fig. S1), the improvements are very limited, as a large area (i.e., western and northeastern United States) has few improvements.

Mean seasonal cycles for soil temperature and moisture are analyzed in Fig. 7. The error bars (two standard deviations) are used to represent observed data uncertainty. The observed and simulated soil temperature averaged over 137 stations for three soil layers show that the mean seasonal cycles in VIC412 and VIC403 are comparable to the observations for all three soil layers (Figs. 7a–c), although VIC412 is closer to the observations in the third soil layer (40–100 cm). For soil moisture evaluations, VIC412 and VIC403 are comparable

FIG. 6. Comparison of long-term mean seasonal cycles calculated from observed/reference and simulated (a) ET, (b) Q , (c) SWNET, (d) LWNET, (e) SH and (f) LH datasets. The envelope between GLEAM and MTE FLUXNET ET is used as an uncertainty estimate for reference ET; error bar is used for Q , SWNET, and LWNET; and yellow shading is an error estimate from Jung et al. (2009). The spatial average is calculated in the southeastern U.S. region ($67^{\circ}\text{--}90^{\circ}\text{W}$, $25^{\circ}\text{--}40^{\circ}\text{N}$).

with the observations for all soil layers and all three regions (Figs. 7d–p), although there are more cases where VIC403 is closer to the observations than VIC412. However, two common features exist in both VIC403 and VIC412. First, there is a very weak seasonal

variability in the top soil layers (0–10 cm) for all three regions (Figs. 7d,j,n); the small bare soil fraction may be a potential cause of this weak variability (Xia et al. 2014). However, it still exists in VIC412 and is not solved yet. This will need the VIC community to make further

FIG. 7. Comparison of long-term mean seasonal cycles calculated from observed and simulated soil temperature (ST; K) and volumetric moisture (SM; unitless): (a) 0–10-cm ST1, (b) 10–40-cm ST2, (c) 40–100-cm ST3, (d) Illinois 0–10-cm SM1, (e) Illinois 10–40-cm SM2, (f) Illinois 40–100-cm SM3, (g) Illinois top 1-m SM, (h) Illinois top 2-m SM, (i) Oklahoma 0–10-cm SM1, (j) Oklahoma 10–40-cm SM2, (k) Oklahoma 40–100-cm SM3, (l) Oklahoma top 1-m SM, (m) West Texas 0–10-cm SM, (n) West Texas 10–40-cm SM, (o) West Texas 40–100 cm SM, and (p) West Texas top 1-m SM.

efforts to fix this issue in the future. Second, the simulated soil moisture in both VIC403 and VIC412 is drier than the observations in the fourth soil layer (1–2 m) in Illinois. In particular, VIC412 is much drier (Fig. 7g). This needs further research on both soil and hydrology physics and parameter aspects in the future.

2) RMSE, TAYLOR SKILL SCORE, AND ANOMALY CORRELATION ANALYSIS

(i) Water budget components

For monthly data, we calculate RMSE, Taylor skill score, and anomaly correlation. As we described in section 3, we use the difference of the metrics to discuss improvement or deterioration for the two model versions. Figure 8 shows dRMSE, dS, and dAC (from left to right) when USGS Q , MTE ET, and GLEAM ET are used as the reference (from top to bottom).

The reduction of Q RMSE in the southeastern United States (Fig. 8a) is attributed to the calibrated soil and hydrologic parameters (Troy et al. 2008). This also leads to a reduction of ET RMSE when two reference ET products are used in the Southeast (Figs. 8d,g). In the Great Plains and Sierra Nevada region, there are also Q RMSE reductions, but for ET RMSE, they are not significant. These conflicting results are due to the error from an imbalance in the reference water budget which will be discussed in section 5b. The reduced RMSE and significantly increased AC (Figs. 8f,i) lead to an increase in S in the Southeast (Figs. 8e,h) for ET but not for Q (Figs. 8b,c). This can be explained in the supplemental text and Fig. S2. The S is related to correlation coefficient R and standard variance ratio SR. When both R and SR are equal to 1, then $S = 1$ and it yields a perfect score. An increase in R always leads to an increase in S . When SR is larger than 1, a decrease (increase) in SR

FIG. 8. Statistics analysis for Q and ET when (a)–(c) USGS Q , (d)–(f) MTE FLUXNET ET, and (g)–(i) GLEAM ET datasets are used as the references. The (left) dRMSE (mm month^{-1}), (center) dS (unitless), and (right) dAC (unitless) are shown when VIC412 and VIC403 are compared. The insignificant dAC at the 95% confidence level is masked out.

leads to an increase (decrease) in S . When SR is smaller than 1, an increase (decrease) in SR leads to an increase (decrease) in S . In the Southeast, there is a small positive R difference and the SR is increased when its value is larger than 1. This results in a small S difference as shown in Fig. 8b. In the western United States, S values are increased due to the fact that R increases and S is closer to 1 when comparing VIC412 with VIC403. When two different reference ET products are used, they may give completely different results. For example, in the northwestern United States, AC decreases significantly when MTE ET is used as the reference; AC increases significantly when GLEAM ET is used as the reference. A similar case can be found in the eastern United States (Figs. 8f,i). Such an evaluation produces conflicting results. Therefore, the ET evaluation may need several reference datasets to obtain a more robust conclusion. Overall, this analysis shows that there are large/significant improvements in the southeastern United States for Q and ET when these three metrics are considered. In other parts of United States, there are small improvements, although AC significantly decreases in the

southern Great Plains for both Q and ET, mainly due to the model upgrade (see section 5).

(ii) Energy components

Figure 9 shows the evaluation for SWNET (first column), LWNET (second column), SH (third column), and LH (fourth column). VIC412 reduces RMSE and significantly increases AC of the SWNET in the western mountainous region (Figs. 9a,i) when compared with VIC403, due to the snowpack processes update discussed in section 4a(2). This upgrade also significantly increases AC values for LWNET (Fig. 9j). For LWNET (Fig. 9f), VIC412 increases S in the northeastern and northwestern United States, but decreases S in the southeastern, north-central, and western United States. The reason for the S decrease is that the correlation coefficient decreases in the Southeast (Fig. S2f) and the standard variance ratio becomes much less than 1 in the north-central and western United States when comparing VIC412 and VIC403 (Figs. S2e, S2g, S2h). This leads to a negative dS value in Fig. 9f, suggesting a skill score decrease.

FIG. 9. As in Fig. 8, but for SWNET, LWNET, SH, and LH, respectively. The (top) dRMSE (W m^{-2}), (middle) dS (unitless), and (bottom) dAC (unitless) are plotted.

For SH and LH, VIC412 largely reduces RMSE (Figs. 9c,d) and increases S (Figs. 9g,h) in the southeastern United States where LH AC is significantly increased, although SH AC is significantly decreased. The spatial pattern of SH and LH dAC is generally opposite for most parts of the United States except the southern Great Plains. This means that the LH simulation skill improvements compensate for the SH simulation skill deterioration so that a seesaw phenomenon can be found in this study (Figs. 9k,l). Therefore, it is still a very challenging issue how to increase both SH and LH AC to improve the VIC model simulation.

(iii) TWSCA

Figure 10 shows the evaluation of monthly TWSCA when GRACE-based data are used as the reference. The results show that there is negative dRMSE (Fig. 10a) and positive dS (Fig. 10b) and dAC (Fig. 10c) in the southeastern United States, suggesting that VIC412 reduces RMSE and increases the simulation skill score, and in particular significantly increases the simulation skill. In the Great Plains, VIC412 increases RMSE

(positive dRMSE) and reduces the simulation skill score (negative dS). As NLDAS-2 P is the same for VIC403 and VIC412, and keeping in mind that $\text{TWSC} = P - (\text{ET} + Q)$, the main errors come from both ET and Q as shown in Fig. S3. The temporal variations of monthly mean observed and simulated TWSCA averages for the western (west of 100°W) and eastern (east of 100°W) United States are shown in Figs. 10d and 10e. The yellow-shaded area represents the uncertainty estimates derived from GRACE data. In the western United States, VIC412 and VIC403 simulations are similar and capture the monthly variation of GRACE-based TWSCA well. However, in the eastern United States the VIC412 simulation is closer to the GRACE-based TWSCA when compared to the VIC403 simulation, which mainly contributes to the RMSE reduction in the southeast.

(iv) Monthly soil temperature, monthly and daily soil moisture

The metrics for monthly soil temperature and moisture are listed in Table 2. For soil temperature, there is

FIG. 10. As in Fig. 8, but for TWSCA [(a) dRMSE (mm month^{-1}), (b) dS (unitless), and (c) dAC (unitless) are shown]. Time series of TWSCA is shown over the (d) western and (e) eastern United States divided by 100°W .

no significant AC change, and the RMSE difference is very small for all three soil layers except the 40–100-cm soil layer, where the model decreases RMSE from 2.08 K in VIC403 to 1.35 K in VIC412. Generally speaking, VIC412 and VIC403 are comparable. RMSE and S changes are mixed for all soil layers and the three regions. However, in Illinois AC is significantly reduced in the 25-cm and 70-cm soil layers, which leads to a significant decrease in the top 1-m and top 2-m soil layers, although AC has a significant increase in the 150-cm soil layer. In Oklahoma, VIC403 and VIC412 are comparable except for a significant AC decrease in the top 10-cm soil layer. In west Texas, VIC412 AC is significantly reduced for all three soil layers except for 10–40 cm, where VIC412 is reduced but not significantly.

To further examine the skill of soil moisture in both VIC403 and VIC412, but over a wider range of climatic regions over all of the CONUS, three additional in situ soil moisture networks were used for evaluation: 1) USDA ARS experimental watersheds (Jackson et al. 2010); 2) the USDA SCAN (Schaefer et al. 2007), and 3) the NOAA USCRN (Diamond et al. 2013; Bell et al.

2013). Both SCAN and USCRN provide hourly measurements at multiple soil depths, while ARS only provides surface 10-cm soil moisture measurements. However, the seven ARS sites are unique, as they represent the average of several nearby individual in situ locations to better represent soil moisture variability within a given model grid cell. Daily averages of the three networks were compared to daily averages of both VIC403 and VIC412 for both the surface and the top 1-m soil moisture. The combination of the deeper soil moisture measurements from SCAN and USCRN to obtain the observed top 1-m soil moisture, the quality control strategy, and other details of the evaluations can be found in Kumar et al. (2018a). The NASA Land Surface Verification Toolkit (LVT; Kumar et al. 2012) was used to perform the evaluations. The AC and the unbiased RMSE (ubRMSE; $\text{m}^3 \text{m}^{-3}$) metrics are shown in Table 3. Consistently between networks, the AC decreases and the ubRMSE increases when going from VIC403 to VIC412, for both surface and top 1-m soil moisture. Generally, the top 1-m soil moisture performs better than the surface soil moisture

TABLE 3. Metrics for surface (SM1/0–10 cm) and top 1-m (0–100 cm) daily soil moisture compared to 7 USDA ARS sites, 119 SCAN sites, and 86 USCRN sites. All values are at the 95% confidence level. AC is the anomaly correlation (bold font indicates significant degradation), and ubRMSE is the unbiased RMSE ($\text{m}^3 \text{m}^{-3}$).

Variable	AC		ubRMSE	
	VIC403	VIC412	VIC403	VIC412
7 USDA ARS sites (2001–16)				
SM1	0.546	0.396	0.044	0.056
119 SCAN quality-controlled sites (2000–14)				
SM1	0.480	0.384	0.063	0.069
Top 1 m	0.591	0.517	0.043	0.047
Variable	VIC403	VIC412	VIC403	VIC412
86 USCRN sites (2010–17)				
SM1	0.443	0.361	0.060	0.066
Top 1 m	0.632	0.543	0.045	0.051

in both versions of VIC. Further analysis found that points with the largest decreases in AC tended to be the southeastern or central United States (Fig. S4). VIC412 had earlier been shown to improve the evaluation of Q and ET in the Southeast over VIC403, primarily due to the calibrated soil and hydrologic parameters when USGS observed streamflow is used. However, the soil moisture skill has decreased, particularly for the surface layer. This suggests that calibration of streamflow cannot ensure that soil moisture simulation is improved. More details can be discussed in section 5f.

5. Discussion

a. Impact of calibrated soil and hydrologic parameters and model version upgrade

In general, the results show that VIC412 improves total runoff and ET simulation in the Southeast and some parts of the western United States when compared with VIC403. The same conclusion can be drawn for the sensible and latent heat flux simulations. However, it remains unclear how much the calibrated soil and hydrologic parameters and model version upgrades respectively contribute to these improvements. To answer this question, we introduce the research NLDAS-2 VIC405 model for this comparison. VIC405 and VIC403 have very similar features for the model but use different soil and hydrologic parameters. In VIC403, the default soil and hydrologic parameters (Mitchell et al. 2004) are used. In VIC405, the calibrated soil and hydrologic parameters (Troy et al. 2008) are used. The difference between these two runs mainly reflects the impact of calibrated parameters, as there are few

differences in model physics/parameterizations (Fig. 1). In VIC412, the same calibrated parameters are used. Therefore, the difference between VIC412 and VIC405 may reflect the impact of the model version upgrade.

Figure 11 shows dRMSE, dS, and dAC (from top to bottom) for Q , ET, SH, and LH (from left to right) between VIC405 and VIC403. In this section, we use 24 years (1984–2007) of monthly data to calculate the metrics, as the research VIC405 data were not extended past 2007. The results show that RMSE of Q is largely reduced in the Sierra Nevada, Great Plains, and Southeast (Fig. 11a), which is consistent with the conclusion in Troy et al. (2008). For ET, SH, and LH (Figs. 11b–d), RMSE reduction appears only in the Southeast, but there is a moderate RMSE increase in the Great Plains and Great Lakes and in some regions of the western United States. This result can be explained based on two factors: 1) as VIC403 is calibrated using observed streamflow, an improvement in runoff simulation does not necessarily improve the ET, SH, and LH simulation and 2) the reference data may have large errors and uncertainties (e.g., water and energy imbalance). For the Taylor skill score S , VIC405 has an increased S value in the Southeast, southern Great Plains, and some regions of the northwest for ET and LH (Figs. 11f,h), but it reduces S values for many regions in the central and western United States. For the SH comparison (Fig. 11g), VIC405 increases S values in the southeast and reduces S values for the northwestern coast, Wisconsin, and Illinois. For the Q comparison (Fig. 11e), VIC405 increases S values for the western United States except for Montana, North Dakota, southern Wyoming, and northern Colorado, which is consistent with VIC412 (Fig. 8b), suggesting calibrated parameters have major impact on the S value increase. VIC405 significantly increases AC values for ET and LH in the eastern United States and decreases AC values for SH in the Southeast. In the Great Plains, VIC405 significantly increases AC values for ET, SH, and LH, but not for Q , which has reduced AC values. In the western United States, VIC405 significantly reduces AC values for ET, SH, and LH, but increases AC values for Q .

Figure 12 shows dRMSE, dS, and dAC for Q , ET, SH, and LH between VIC412 and VIC405. In part of the southern Great Plains, VIC412 increases RMSE (reduces S) for all four variables examined. However, RMSE is reduced (S is increased) for ET, SH, and LH in most regions of the United States (Figs. 12a–h). In the AC comparison, VIC412 has mixed results for Q in the Great Plains as it significantly increases AC in one place and significantly reduces AC in another (Fig. 12i).

FIG. 11. As in Q and ET in Fig. 8 and SH and LH in Fig. 9, but when VIC405 and VIC403 are compared (VIC405-VIC403). This comparison displays impact of calibrated soil and hydrological parameters on VIC model performance.

However, VIC412 significantly increases AC for ET and for SH over the United States except for the Great Plains. If we examine three metrics and four variables, VIC412 performance is improved by reducing RMSE and increasing S and AC in many regions of the United States, suggesting that the model upgrade is moving in the correct direction. The Great Plains region is an exception as VIC412 performance is degraded when compared with VIC405. The reason remains unclear and needs further investigation from the VIC model development community.

When we compare Figs. 11 and 12, we find that in the southeast both calibrated parameters and model versions have positive impact on VIC model improvement by reducing RMSE for Q , ET, SH, and LH and increasing S and AC for ET, SH, and LH. This leads to significant improvement in the southeast when comparing VIC412 and VIC403 (Fig. 8), although calibration of model parameters in VIC412 is still needed. In the Great Plains, the impact of the calibrated parameters is positive and impact of the model upgrade is negative so that there is no improvement or even a

deterioration (AC reduction) when comparing VIC412 with VIC403, suggesting a recalibration of VIC412 is needed to improve its performance. This work is being planned between two NLDAS teams and the VIC model development community (e.g., University of Washington and Princeton University). In the western United States, calibrated parameters and the model version upgrade also have an opposite impact and lead to a compromised result (e.g., calibrated parameters have positive impacts and model version upgrade has negative impact or vice versa) as shown in Fig. 8. Therefore, when a model has some major upgrades, a recalibration may be a key step to improve model performance.

b. Water and energy budget imbalance issue from reference data

Although we use either reference product uncertainty estimates or error bars to quantify reference data uncertainties when we evaluate our model simulations, nonclosure error in the water and energy budget is still an important issue (Gao et al. 2010; Sahoo et al. 2011; Zhang et al. 2018). Even for the Mississippi River basin,

FIG. 12. As in Fig. 11, but when VIC412 and VIC405 are compared (VIC412-VIC405). This comparison displays the impact of model version upgrade on VIC model performance.

nonclosure errors calculated from multiple remote sensing estimates are larger than the observed streamflow, even after these estimates are bias corrected (Sheffield et al. 2009). As our reference data come from various sources such as USGS total runoff derived from observations and gridded MTE ET data, large imbalance errors exist in the water budget (Fig. 13a). In general, in the eastern United States and Great Plains, the water imbalance is less than 10% of mean annual precipitation. However, in the western United States, the water imbalance reaches more than 30% of mean precipitation in many regions. These errors are comparable with those (5%–25% of mean annual precipitation for multiple global river basins) found in Sahoo et al. (2011), as there may be larger errors for grid points than for the basin average. This may be a partial reason for the conflicting information available for evaluating different variables, as the water balance is not preserved between the various evaluation reference datasets.

As multiple data sources are also used for energy budget component evaluation, the nonclosure errors for the energy budget also exist in such an analysis.

Figure 13b shows the relative error (bias) when energy nonclosure error is calculated as $[100 \times (RNET - SH - LH)/RNET]$. The largest error is in the southwestern region; there is moderate error along the western coast, the Great Plains except for Texas, the central-north including the Great Lakes, and parts of the Southeast and Northeast; and there are relatively small errors in the other regions. The cause may be that there is lower net radiation (low SWNET and high LWNET; see Figs. 4a,b) when compared with multimodel NLDAS-2 products, in particular during the summer season (Xia et al. 2016a,b). As there are no long-term monthly observations from the towers, it is very difficult to further examine the reference net radiation accuracy.

Therefore, both water and energy nonclosure errors can affect our evaluation work presented here. For the water component evaluation, in the central and eastern United States the effect may be relatively small and evaluation results are more reliable. However, in the western United States, in mountainous regions in particular, model evaluation results should be interpreted

FIG. 13. Relative imbalance error of the (a) water and (b) energy budget. It is relative to mean annual precipitation for the water budget and mean annual net radiation for the energy budget (NLDAS-2 precipitation, USGS total runoff, and MTE ET are used to calculate water budget imbalance error; the GEWEX net radiation and MTE sensible and latent heat flux are used to calculate energy budget imbalance error).

cautiously. For the energy component evaluation, in most parts of the United States except for the Southwest, Great Lakes, and Northeast, the effect of non-closure error may be small to moderate and the evaluation results are relatively reliable. For basin-scale work, Pan et al. (2012) have developed a multisource constraint balance method to generate a water-balanced dataset with various water budget components. This method has been extended from the basin-scale to gridpoint scale at 0.5° resolution and monthly scale over the globe (Zhang et al. 2018). Further study is underway to provide more refined multisource and balance-constrained water budget analysis at higher resolution over United States, and we can reevaluate the model simulations once that becomes available.

c. *In situ* observed data quality and spatial scale mismatch issue

Generally, in situ observed data quality is limited by short periods of record, many missing records, and limited numbers of stations or eddy flux tower sites. For example, the North American Soil Moisture Database collected data from over 1800 stations over the United States (Quiring et al. 2016). The measurements are

largely missing in winter and for deep soil layers (>60 cm) due to sensor failures. Although many records have been quality controlled by the network developers, during the cold season (October–April), many measurements with the sensor malfunction problem (when the soil is partially frozen, a sensor can measure only the liquid moisture rather than total soil moisture) still exist (Xia et al. 2015c). This is a major reason why we select four regions in this study, as they have relatively complete records with a quality control procedure (Xia et al. 2015c). A lot of missing records also exist in the soil temperature data, as indicated by Xia et al. (2013). Even though data have been carefully quality controlled, a 3%–6% measurement error for soil moisture and a 0.5°C error for soil temperature still exist.

In this study, we use a simple spatial averaging method to evaluate model simulations against in situ observations. This method has been used to evaluate model products in the first phase (Robock et al. 2003) and second phase (Xia et al. 2013, 2015a,b) of the NLDAS system. A recent study from Dirmeyer et al. (2016) shows that there is no significant difference between AC calculated using spatially averaged data or using individual station data. However, it has a large effect on the calculation of errors and skill scores, as errors or biases can cancel each other when multiple stations are used. The spatial scale mismatch issue has been a challenge as many observed data such as in situ soil moisture have a small scale (~ 10 – 100 m) related to hydrological processes and soil and vegetation characteristics. This characteristic small scale also exists in measured fluxes from eddy flux towers. Typical networks include the AmeriFlux network of FLUXNET (Baldocchi et al. 2001) and the Atmospheric Radiation Measurement/Cloud and Radiation Testbed (ARM/CART; Robock et al. 2003). In general, these flux data are representative of the area within about 50–100 m of the towers, as these flux data are closely related to boundary layer processes and soil and vegetation characteristics (Kustas et al. 2004; McCabe and Wood 2006; Li et al. 2008). However, such a high-resolution (i.e., ~ 100 m) system is under development and is less mature (Chaney et al. 2016; Cai et al. 2017) over the NLDAS domain. Furthermore, the computational burden of running the land surface model at such a resolution for a more than a 30-yr period is currently challenging for the continental United States. Therefore, the simple spatial averaging discussed earlier represents a necessary compromise to compare NLDAS-2 products with in situ measurements from stations or eddy flux towers. Although eddy tower flux data are not used in this study for monthly data evaluation, due to short periods of record and many missing records, they will be used for daily and hourly product evaluation in the future.

It should be noted that there are also scale mismatch errors for gridded products due to different spatial resolutions when the coarse- or fine-resolution reference products are regridded to a 0.125° NLDAS resolution. Such a remapped method has been used in many previous studies (Artan et al. 2013; Velpuri et al. 2013; Cai et al. 2014; Kumar et al. 2014; Xia et al. 2016a,b; Kumar et al. 2017) as the regridded errors are assumed to be small when compared with those differences resulting from the model version upgrade and calibrated model parameters. This approach should be reasonable in this study as we used the same references to compare metrics differences (e.g., dRMSE, dS, dAC) to examine model relative improvement or deterioration.

d. Large negative Rbias in SWE simulations

The large negative Rbias values shown in Fig. 5 may be due to several factors. Pan et al. (2003) has indicated that small simulated SWE values [compared with Snowpack Telemetry (SNOTEL)-observed SWE] are due to less NLDAS-based gauge precipitation when compared with SNOTEL-observed precipitation. As most rain gauge stations are located in valleys rather than on the tops of mountains, gauge-based precipitation is not representative for higher elevations. Although NLDAS-2 precipitation is bias corrected by monthly PRISM precipitation (Daly et al. 1994; Xia et al. 2012a), the precipitation values may be still smaller than SNOTEL observations. Additionally, Royer and Poirier (2010) compared North American Regional Reanalysis (NARR) air temperatures with in situ observations and satellite retrievals and found that the NARR air temperature can be at times 1°–2°C higher. The NLDAS-2 air temperature is downscaled from the NARR product and thus it has the same characteristics. When snowfall is partitioned from total precipitation using 0°C air temperature threshold for VIC403, warm air temperatures will partition precipitation into less snowfall. Furthermore, using the traditional 0°C air temperature threshold to separate total precipitation into liquid (rainfall) and solid (snowfall and frozen rain) parts does result in less snowfall than using the Jordan (1991) algorithm (Xia et al. 2017). VIC412 uses a temperature range (0°–1.5°C) to divide the total precipitation into liquid and solid form. The snowfall and rainfall are proportional to the air temperature when it is within this temperature range. As a result, these three factors lead to less snowfall in the VIC model.

Additionally, this underestimation may also be due to SNODAS overestimating field SWE measurements (Brennan et al. 2013), regridding errors from 1 km to 0.125° spatial resolution, and some unclear weakness related to snowpack physical processes and parameters

in the VIC model, even though VIC412 snowpack-related physical processes have been upgraded (see Fig. 1) and calibrated soil and hydrologic parameters are used. These examinations and sensitivity tests will be performed in the near future in the NLDAS Science Testbed created by the NASA GFSC NLDAS team using the NASA-developed LVT (Kumar et al. 2012).

e. Challenging issues of parameter calibrations in the Great Plains

As analyzed in section 4b, the model version upgrade and parameter calibration have opposite impacts on model performance. In particular, calibrated parameters do increase AC values in the Great Plains and reduce RMSE in the southern Great Plains for ET, SH, and LH simulations. Therefore, a recalibration for VIC412 is an important step to enhance its performance. However, there are three challenges that limit such a calibration. The first is the lack of in situ observations in this region. The observation usually used for VIC model calibration is USGS streamflow, even though there are few streamflow gauges there. Troy et al. (2008) calibrated VIC403 and found some moderate bias reduction in the Great Plains. Recently Newman et al. (2017) calibrated VIC4.1.2.i using a comprehensive methodology and observed streamflow, and the results show very limited improvement for that region. In addition, there are very few long-term (e.g., 10–20 years) in situ ET, SH, LH, soil moisture, and temperature observations in that region. The second challenge is the lack of scientific understanding of land surface physical processes including soil and hydrology physical process (e.g., groundwater dynamics) in arid/semiarid regions where sand is a dominant soil type. For example, when precipitation is partitioned into ET and total runoff in arid and semiarid regions, most of the precipitation becomes ET and only a small part of the precipitation goes into total runoff. A 10%–20% error in precipitation can lead to 100% errors in runoff simulations (Milly and Dunne 2002). Therefore, even when soil and hydrologic parameters are calibrated to match the model against USGS-observed streamflow, these parameters do not necessarily improve the ET simulation as the latter has the larger share. Even after a strict calibration, the VIC model still produces a large runoff error when compared with USGS observations in arid and semiarid regions such as the Great Plains and some regions of the western United States (Newman et al. 2017). The contributing factors may be the lack of some physical processes such as groundwater and irrigation, as well as an inappropriate partitioning of evapotranspiration into bare soil evaporation, canopy evaporation, and transpiration (Bohn and

Vivoni 2016). The third factor is how to select the calibration reference datasets. In the Great Plains, though there are few in situ observations, there are quite a lot of reference datasets such as remote sensing ET data [e.g., MODIS (Mu et al. 2011), ALEXI (Anderson et al. 2011), Surface Energy Balance System (SEBS; Su 2002), and GLEAM (Miralles et al. 2011)], FLUXNET-based gridded ET (Jung et al. 2009), and other reanalysis ET products. In addition, FLUXNET-based gridded sensible heat, remotely sensed land surface temperature, soil moisture, terrestrial water storage (e.g., GRACE), and radiation fluxes also exist. They are good alternatives to use to calibrate the VIC model soil and hydrologic parameters. However, as these reference products have their own errors and uncertainties, determining how to use them to calibrate the VIC models is a challenge. Conducting an error estimate of these reference data is the first step, as many products do not have their own error and uncertainty estimates. Second, the target variables selected for model calibration should be a major component in the water and energy cycle and have a strong relationship with other major water and energy budget components.

f. Calibration of single variable versus multiple variables

Calibration of model parameters is a very important procedure for land surface or hydrological models to ensure good skills in predicting many physical processes. Furthermore, many model parameters are not directly measurable and need to be calibrated against observed data. For the VIC model, Troy et al. (2008) used the observed streamflow interpolated from 1130 USGS small basins to calibrate VIC403 in every $1^\circ \times 1^\circ$ box to reduce the large positive bias in simulated total runoff in southeastern United States. The results are very encouraging as calibrated parameters in VIC405 and VIC412 largely reduce total runoff and ET bias in that region, compared with default soil and hydrologic parameters (Maurer et al. 2002). Unfortunately, calibrated parameters are deteriorating soil moisture simulations for both surface and top 1-m soil layers as the AC values decrease and RMSE errors increase. Therefore, when the error in the calibration target variable is reduced, errors in other variables may be increased. This is a major motivation to extend single-criterion calibration (Duan et al. 1994) to multicriteria calibration (Yapo et al. 1998). It will be expected that using both observed streamflow and soil moisture to calibrate soil and hydrologic parameters may potentially improve overall model performance, although scale mismatch will be an issue for soil moisture calibration. Although such a calibration cannot obtain the best parameter set for all

calibration target variables, it can obtain a “pareto set,” which is optimal for model simulation (Xia et al. 2002). In addition, model calibration may be also model version dependent. Thus, a recalibration for VIC412 using multiple variables (e.g., streamflow, ET, soil moisture, sensible heat flux, SWE, land skin temperature) is recommended for the modeling community. Efforts in improving calibration density, for example, using more catchments (Oubeidillah et al. 2014) or through parameter regionalization (Mizukami et al. 2017), also help refine model skills. From a soil physics viewpoint, we need to carefully verify soil texture, hydrologic conductivity, wilting point, field capacity, and other soil hydrologic parameters for both in situ sites and nearest grid cells in the VIC model. From a science viewpoint, we need to study the relationships between soil moisture, soil and hydrologic parameters, vegetation parameters, and evaporation and transpiration to enhance science understanding. Through all these efforts, we may be able to improve the overall performance of the VIC model for those variables closely related to drought monitoring and forecast work. When multiple water and energy fluxes are used for calibration purposes, the imbalance of water and energy budget needs to be assessed first and then the balance-corrected fluxes could be used for model calibration.

The highly efficient calibration method (Duan et al. 1994; Yapo et al. 1998; Jackson et al. 2003), high performance computer, and parallelized model code are important tools to speed up the multivariable and multimetrics calibration process when long-term gridded reference data are used for CONUS. The VIC412 simulation was performed using the NASA LIS software framework in a parallel mode to greatly speed up its simulation process compared to VIC403 and VIC405. A planned recalibration and testing of different model options for VIC412 using both USGS streamflow/runoff and ensemble mean ET from multiple sources is underway.

g. Role of physical processes and implications of the evaluation results on drought monitoring

Although we address the importance of calibrations in this study, the model parameters calibration is not a physically based model development procedure. As mentioned in section 5f, the calibration might improve the simulation of some variables, but it might also deteriorate others as the calibration is a nonphysical approach, which is largely dependent on model versions and on the quality of the reference data. Therefore, for the model developer, a more physically based approach is to enhance the science understanding of model physical processes. For example, the VIC model is

lacking irrigation process and groundwater dynamics, which are two important physical processes for water cycle simulation, particularly in the Great Plains. The lack of such processes might be directly related to the large simulation errors that appeared in the Great Plains. In addition, for total evapotranspiration calculation, VIC might partition more water vapor into transpiration (Bohn and Vivoni 2016; Kumar et al. 2018b). The less evaporation in bare soil has led to smaller seasonal variability in top 10-cm soil moisture (Xia et al. 2014). Therefore, it is a critical step to develop and/or add physically based parameterization schemes to the VIC model in order to improve overall model performances. In summary, enhancement to model physical processes, improvement of observed and reference data quality, and appropriate calibration of model parameters need to work together to comprehensively improve model performance.

There are many applications using NLDAS products and LIS infrastructure for drought monitoring. Drought can be generally classified as meteorological drought, agricultural drought, and hydrological drought. Broadly speaking, precipitation is generally used to monitor meteorological drought, soil moisture and ET are used to monitor agricultural drought, and total runoff/streamflow is used to monitor hydrological drought. As presented in this study, improved runoff and ET simulation will enhance flash agricultural and hydrological drought monitoring capability. However, VIC412 will deteriorate general agricultural drought monitoring ability as soil moisture simulations are becoming worse for all soil layers. Given the relatively large amount of soil moisture observations, soil moisture simulations can be improved when the VIC model is calibrated using both soil moisture and streamflow observations. Furthermore, improvements of soil and hydrology physics in the VIC model as well as the addition of irrigation and groundwater procedures have the potential improve soil moisture simulation accuracy.

h. Future work

In this study, we used monthly reference products, including station and grid-based data. These products are either measured from stations or derived from satellite remote sensing and reanalysis products. It is very well known that VIC can simulate multiple temporal and spatial scale land surface physical processes (Liang et al. 1994; Wood et al. 1997, 1998; Nijssen et al. 2001, 2003; Maurer et al. 2002; Cherkauer et al. 2003; Sheffield and Wood 2007). The temporal time scales are hourly, daily, monthly, annual, and decadal. The spatial scale covers a single grid point, watershed, large basin, continent, and the globe. As Mitchell et al. (2004) indicated,

evaluating the monthly diurnal cycle for radiation fluxes, SH, LH, and land skin temperature is important for understanding land surface model physical processes, which are largely controlled by solar radiation. In addition, daily variability is also important for some variables such as streamflow, soil moisture, and ET, as they are usually used for daily to weekly drought and flood monitoring.

In the future, a more comprehensive evaluation and comparison against daily deep soil layer soil moisture, soil temperature, ET, streamflow, and SWE is needed using all available daily in situ observations and remote sensing data. A further evaluation at an hourly time scale such as the diurnal cycle analysis presented in Robock et al. (2003) will be useful for VIC model development when radiation fluxes, sensible heat flux, latent heat flux, ground heat flux, land skin temperature are measured from towers in various networks such as ARM/CART and AmeriFlux. These evaluations will be planned in the future.

When we extend our evaluations from monthly to daily and hourly time scales, we also need to diagnose why the model upgrade leads to deteriorated performance in the Great Plains, for example, and what physics upgrade or bug fix leads to this deterioration. For operational implementation and application purposes, we also need to recalibrate the VIC412 model using an optimal method when multiple variables and metrics are considered. Therefore, we encourage the VIC model development community in governmental agencies and academia to investigate these issues in the future. The efforts of the community will help our NLDAS teams speed up VIC model upgrades and its operational implementation.

6. Summary and conclusions

In this study, we used multiple variables (i.e., water fluxes, energy fluxes, state variables) and metrics (e.g., bias, Rbias, RMSE, S , and AC) with the significance test for monthly time scales to evaluate the VIC model. The results show that for most evaluated variables, the latest VIC model version VIC412 has a large and significant improvement in the southeastern United States when compared with current operational NLDAS-2 VIC403. There is less improvement in other regions. In situ soil temperature evaluation from 137 stations shows that the VIC412 and VIC403 simulations are comparable. However, the in situ soil moisture evaluation shows that VIC412 is degrading model performance for some layers/locations when compared with VIC403. This conclusion should be interpreted cautiously because only a limited number of soil moisture stations are used.

The model version upgrade and parameter calibration have an opposite impact on VIC model performance in the Great Plains: calibrated parameters improve model performance and the model upgrade degrades model performance, suggesting a recalibration of VIC412 is needed. In addition, due to the nonclosure errors of water and energy budget, there are some conflicting results. Therefore, a cautious interpretation should be taken from this study.

It will be also extended from monthly time scales to daily and hourly time scales as the current temporal aggregation to monthly values limits the applicability of the framework to study issues such as the diurnal cycle and daily variations. In addition, water and energy budget imbalance errors bring some conflicting results so that it is difficult to draw a consistent conclusion for some regions or variables. With the extension of the constraint balance error approach developed by Pan et al. (2012) from river basin scale to finescale grid point (Zhang et al. 2018), a reevaluation will be performed in the future.

Acknowledgments. This research is supported by NOAA Climate Program of Office MAPP (Model, Analysis, Predictions, and Projection) project. The authors thank Mary Hart at EMC, who edited the first draft of this manuscript; Anil Kumar and Jacob Carley at EMC made the EMC internal review, and their comments greatly improved quality of this manuscript. Computing for the VIC412 simulation using NASA LIS was supported by the resources at the NASA Center for Climate Simulation.

REFERENCES

- Anderson, M. C., C. Hain, B. Wardlow, A. Pimstein, J. Mecikalski, and W. P. Kustas, 2011: Evaluation of drought indices based on thermal remote sensing of evapotranspiration over the continental United States. *J. Climate*, **24**, 2025–2044, <https://doi.org/10.1175/2010JCLI3812.1>.
- Artan, G. A., J. P. Verdin, and R. Lietzow, 2013: Large scale snow water equivalent status monitoring: Comparison of different snow water products in the upper Colorado basin. *Hydrol. Earth Syst. Sci.*, **17**, 5127–5139, <https://doi.org/10.5194/hess-17-5127-2013>.
- Baldocchi, D., and Coauthors, 2001: FLUXNET: A new tool to study the temporal and spatial variability of eco-system-scale carbon dioxide, water vapor, and energy flux densities. *Bull. Amer. Meteor. Soc.*, **82**, 2415–2434, [https://doi.org/10.1175/1520-0477\(2001\)082<2415:FANTTS>2.3.CO;2](https://doi.org/10.1175/1520-0477(2001)082<2415:FANTTS>2.3.CO;2).
- Bell, J., and Coauthors, 2013: U.S. Climate Reference Network soil moisture and temperature observations. *J. Hydrometeorol.*, **14**, 977–988, <https://doi.org/10.1175/JHM-D-12-0146.1>.
- Bohn, T. J., and E. R. Vivoni, 2016: Process-based characterization of evapotranspiration sources over the North American monsoon region. *Water Resour. Res.*, **52**, 358–384, <https://doi.org/10.1002/2015WR017934>.
- Brennan, A., P. C. Cross, M. Higgs, J. P. Beckmann, P. W. Klaver, B. M. Scurlock, and S. Creel, 2013: Inferential consequences of modeling rather than measuring snow accumulation in studies of animal ecology. *Ecol. Appl.*, **23**, 643–653, <https://doi.org/10.1890/12-0959.1>.
- Cai, X., Z.-L. Yang, Y. Xia, M. Huang, H. Wei, R. Leung, and M. B. Ek, 2014: Assessment of simulated water balance from Noah, Noah-MP, CLM, and VIC over CONUS using the NLDAS testbed. *J. Geophys. Res. Atmos.*, **119**, 13 751–13 770, <https://doi.org/10.1002/2014JD022113>.
- , and Coauthors, 2017: Validation of SMAP soil moisture for the SMAPVEX15 field campaign using a hyper-resolution model. *Water Resour. Res.*, **53**, 3013–3028, <https://doi.org/10.1002/2016WR019967>.
- Chaney, N. W., P. Metcalfe, and E. F. Wood, 2016: HydroBlocks: A field-scale resolving land surface model for application over continental extents. *Hydrol. Processes*, **30**, 3543–3559, <https://doi.org/10.1002/hyp.10891>.
- Cherkauer, K. A., L. C. Bowling, and D. P. Lettenmaier, 2003: Variable Infiltration Capacity (VIC) cold land process model updates. *Global Planet. Change*, **38**, 151–159, [https://doi.org/10.1016/S0921-8181\(03\)00025-0](https://doi.org/10.1016/S0921-8181(03)00025-0).
- Clow, D. W., L. Nanus, K. L. Verdin, and J. Schmidt, 2012: Evaluation of SNODAS snow depth and snow water equivalent estimates for the Colorado Rocky Mountains, USA. *Hydrol. Processes*, **26**, 2583–2591, <https://doi.org/10.1002/hyp.9385>.
- Cosgrove, B. A., and Coauthors, 2003: Land surface model spin-up behavior in the North American Land Data Assimilation System (NLDAS). *J. Geophys. Res.*, **108**, 8845, <https://doi.org/10.1029/2002JD003316>.
- Daly, C., R. P. Neilson, and D. L. Phillips, 1994: A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *J. Appl. Meteor.*, **33**, 140–158, [https://doi.org/10.1175/1520-0450\(1994\)033<0140:ASTMFM>2.0.CO;2](https://doi.org/10.1175/1520-0450(1994)033<0140:ASTMFM>2.0.CO;2).
- Diamond, H., and Coauthors, 2013: U.S. Climate Reference Network after one decade of operations: Status and assessment. *Bull. Amer. Meteor. Soc.*, **94**, 485–498, <https://doi.org/10.1175/BAMS-D-12-00170.1>.
- Dirmeyer, P. A., and Coauthors, 2016: Confronting weather and climate models with observational data from soil moisture networks over the United States. *J. Hydrometeorol.*, **17**, 1049–1067, <https://doi.org/10.1175/JHM-D-15-0196.1>.
- Duan, Q., S. Sorooshian, and V. K. Gupta, 1994: Optimal use of the SCE-UA global optimization method for calibrating watershed models. *J. Hydrol.*, **158**, 265–284, [https://doi.org/10.1016/0022-1694\(94\)90057-4](https://doi.org/10.1016/0022-1694(94)90057-4).
- Ek, M. B., K. E. Mitchell, Y. Lin, E. Rodgers, P. Grunman, V. Koren, G. Gayno, and J. D. Tarpley, 2003: Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. *J. Geophys. Res.*, **108**, 8851, <https://doi.org/10.1029/2002JD003296>.
- , and Coauthors, 2017: Next phase of the NCEP Unified Land Data Assimilation System (NULDAS): Vision, requirements, and implementation. NLDAS White Paper, 17 pp., http://www.emc.ncep.noaa.gov/mmb/nldas/White_Paper_for_Next_Phase_LDAS_final.pdf.
- Gao, H., Q. Tang, C. R. Ferguson, E. F. Wood, and D. P. Lettenmaier, 2009: Estimating the water budget of major US river basins via remote sensing. *Int. J. Remote Sens.*, **31**, 3955–3978, <https://doi.org/10.1080/01431161.2010.483488>.
- , and Coauthors, 2010: Water budget record from variable infiltration capacity (VIC) model. Algorithm Theoretical Basis Document for Terrestrial Water Cycle Data Records, Algorithm Theoretical Basis Doc., 120–173, <http://hydrology>.

- princeton.edu/~mpan/academics/uploads/content/articles/Water_Cycle_MEaSURES_ATBD_Combined_v1.0.pdf.
- Jackson, C., Y. Xia, M. K. Sen, and P. L. Stoffa, 2003: Optimal parameter and uncertainty estimation of a land surface model: A case study using data from Cabauw, Netherlands. *J. Geophys. Res.*, **108**, 4583, <https://doi.org/10.1029/2002JD002991>.
- Jackson, T., and Coauthors, 2010: Validation of Advanced Microwave Scanning Radiometer soil moisture products. *IEEE Trans. Geosci. Remote Sens.*, **48**, 4256–4272, <https://doi.org/10.1109/TGRS.2010.2051035>.
- Jiménez, C., and Coauthors, 2011: Global intercomparison of 12 land surface heat flux estimates. *J. Geophys. Res.*, **116**, D02102, <https://doi.org/10.1029/2010JD014545>.
- Jordan, R., 1991: A one-dimensional temperature model for a snow cover: Technical documentation for SNTERERM.89. Special Rep. 91-16, Cold Region Research and Engineers Laboratory, U.S. Army Corps of Engineers, Hanover, NH, 61 pp.
- Jung, M., M. Reichstein, and A. Bondeau, 2009: Towards global empirical upscaling of FLUXNET eddy covariance observations: Validation of a model tree ensemble approach using a biosphere model. *Biogeosciences*, **6**, 2001–2013, <https://doi.org/10.5194/bg-6-2001-2009>.
- , and Coauthors, 2011: Recent decline in the global land evapotranspiration trend due to limited moisture supply. *Nature*, **467**, 951–954, <https://doi.org/10.1038/nature09396>.
- Koster, R. D., M. J. Suarez, A. Ducharne, M. Stieglitz, and P. Kumar, 2000: A catchment-based approach to modeling land surface processes in a general circulation model: 1. Model structure. *J. Geophys. Res.*, **105**, 24 809–24 822, <https://doi.org/10.1029/2000JD900327>.
- Kumar, S. V., and Coauthors, 2006: Land information system: An interoperable framework for high resolution land surface modeling. *Environ. Modell. Software*, **21**, 1402–1415, <https://doi.org/10.1016/j.envsoft.2005.07.004>.
- , C. D. Peters-Lidard, J. Santanello, K. Harrison, Y. Liu, and M. Shaw, 2012: Land surface Verification Toolkit (LVT)—A generalized framework for land surface model evaluation. *Geosci. Model Dev.*, **5**, 869–886, <https://doi.org/10.5194/gmd-5-869-2012>.
- , and Coauthors, 2014: Assimilation of remotely sensed soil moisture and snow depth retrievals for drought estimation. *J. Hydrometeorol.*, **15**, 2446–2469, <https://doi.org/10.1175/JHM-D-13-0132.1>.
- , S. Wang, D. M. Mocko, C. D. Peters-Lidard, and Y. Xia, 2017: Similarity assessment of land surface model outputs in the North American Land Data Assimilation System. *Water Resour. Res.*, **53**, 8941–8965, <https://doi.org/10.1002/2017WR020635>.
- Kumar, S., M. Jasinski, D. Mocko, M. Rodell, J. Borak, B. Li, H. K. Beaudoin, and C. Peters-Lidard, 2018a: NCA-LDAS land analysis: Development and performance of a multisensor, multivariate land data assimilation system for the National Climate Assessment. *J. Hydrometeorol.*, <https://doi.org/10.1175/JHM-D-17-0125.1>, in press.
- , T. Holmes, D. M. Mocko, S. Wang, and C. Peters-Lidard, 2018b: Attribution of flux partitioning variations between land surface models over the continental U.S. *Remote Sens.*, **10**, 751, <https://doi.org/10.3390/rs10050751>.
- Kustas, W. P., F. Li, T. J. Jackson, J. H. Prueger, J. I. MacPherson, and M. Wolde, 2004: Effects of remote sensing pixel resolution on modeled energy flux variability of croplands in Iowa. *Remote Sens. Environ.*, **92**, 535–547, <https://doi.org/10.1016/j.rse.2004.02.020>.
- Landerer, F. W., and S. C. Swenson, 2012: Accuracy of scaled GRACE terrestrial water storage estimates. *Water Resour. Res.*, **48**, W04531, <https://doi.org/10.1029/2011WR011453>.
- Li, F., W. P. Kustas, M. C. Anderson, J. H. Prueger, and R. L. Scott, 2008: Effect of remote sensing spatial resolution on interpreting tower-based flux observations. *Remote Sens. Environ.*, **112**, 337–349, <https://doi.org/10.1016/j.rse.2006.11.032>.
- Liang, X., D. P. Lettenmaier, E. F. Wood, and S. J. Burges, 1994: A simple hydrologically based model of land surface water and energy fluxes for GCMs. *J. Geophys. Res.*, **99**, 14 415–14 428, <https://doi.org/10.1029/94JD00483>.
- Lohmann, D., and Coauthors, 2004: Streamflow and water balance intercomparisons of four land surface models in the North American Land Data Assimilation System project. *J. Geophys. Res.*, **109**, D07S91, <https://doi.org/10.1029/2003JD003517>.
- Martens, B., and Coauthors, 2017: GLEAM v3: Satellite-based land evaporation and root-zone soil moisture. *Geosci. Model Dev.*, **10**, 1903–1925, <https://doi.org/10.5194/gmd-10-1903-2017>.
- Maurer, E. P., A. W. Wood, J. C. Adam, D. P. Lettenmaier, and B. Nijssen, 2002: A long-term hydrologically-based data set of land surface fluxes and states for the conterminous United States. *J. Climate*, **15**, 3237–3251, [https://doi.org/10.1175/1520-0442\(2002\)015<3237:ALTHBD>2.0.CO;2](https://doi.org/10.1175/1520-0442(2002)015<3237:ALTHBD>2.0.CO;2).
- McCabe, M. F., and E. F. Wood, 2006: Scale influences on the remote estimation of evapotranspiration using multiple satellite sensors. *Remote Sens. Environ.*, **105**, 271–285, <https://doi.org/10.1016/j.rse.2006.07.006>.
- Michel, D., and Coauthors, 2016: The WACMOS-ET project - Part 1: Tower-scale evaluation of four remote sensing-based evapotranspiration algorithm. *Hydrol. Earth Syst. Sci.*, **20**, 803–822, <https://doi.org/10.5194/hess-20-803-2016>.
- Milly, P. C. D., and K. A. Dunne, 2002: Macroscale water fluxes, 1. Quantifying errors in the estimation of basin mean precipitation. *Water Resour. Res.*, **38**, 1205, <https://doi.org/10.1029/2001WR000759>.
- Miralles, D. G., T. R. H. Holmes, R. A. M. de Jeu, J. H. Gash, A. G. Meesters, and A. J. Dolman, 2011: Global land-surface evaporation estimated from satellite-based observations. *Hydrol. Earth Syst. Sci.*, **15**, 453–469, <https://doi.org/10.5194/hess-15-453-2011>.
- , and Coauthors, 2016: The WACMOS-ET project – Part 2: Evaluation of global terrestrial evaporation data sets. *Hydrol. Earth Syst. Sci.*, **20**, 823–842, <https://doi.org/10.5194/hess-20-823-2016>.
- Mitchell, K. E., and Coauthors, 2004: The multi-institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system. *J. Geophys. Res.*, **109**, D07S90, <https://doi.org/10.1029/2003JD003823>.
- Mizukami, N., M. P. Clark, A. J. Newman, A. W. Wood, E. Gutmann, B. Nijssen, O. Rakovec, and L. Samaniego, 2017: Towards seamless large domain parameter estimation for hydrologic models. *Water Resour. Res.*, **53**, 8020–8040, <https://doi.org/10.1002/2017WR020401>.
- Mu, Q., M. Zhao, and S. W. Running, 2011: Improvements to a MODIS global terrestrial evapotranspiration algorithm. *Remote Sens. Environ.*, **115**, 1781–1800, <https://doi.org/10.1016/j.rse.2011.02.019>.
- Mueller, B., and Coauthors, 2011: Evaluation of global observations-based evapotranspiration datasets and IPCC AR4 simulations. *Geophys. Res. Lett.*, **38**, L06402, <https://doi.org/10.1029/2010GL046230>.

- Nash, J. E., and J. V. Sutcliffe, 1970: River flow forecasting through conceptual model: Part A – A discussion of principles. *J. Hydrol.*, **10**, 282–290, [https://doi.org/10.1016/0022-1694\(70\)90255-6](https://doi.org/10.1016/0022-1694(70)90255-6).
- Newman, A. J., N. Mizukami, M. P. Clark, A. W. Wood, B. Nijssen, and G. Nearing, 2017: Benchmarking of a physically based hydrologic model. *J. Hydrometeorol.*, **18**, 2215–2225, <https://doi.org/10.1175/JHM-D-16-0284.1>.
- Nijssen, B., G. M. O'Donnell, D. P. Lettenmaier, D. Lohmann, and E. F. Wood, 2001: Predicting the discharge of global rivers. *J. Climate*, **14**, 3307–3323, [https://doi.org/10.1175/1520-0442\(2001\)014<3307:PTDOGR>2.0.CO;2](https://doi.org/10.1175/1520-0442(2001)014<3307:PTDOGR>2.0.CO;2).
- , and Coauthors, 2003: Simulation of high latitude hydrological processes in the Torne–Kalix basin: PILPS Phase 2(e): 2: Comparison of model results with observations. *Global Planet. Change*, **38**, 31–54, [https://doi.org/10.1016/S0921-8181\(03\)00004-3](https://doi.org/10.1016/S0921-8181(03)00004-3).
- Niu, G.-Y., and Coauthors, 2011: The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *J. Geophys. Res.*, **116**, D12109, <https://doi.org/10.1029/2010JD015139>.
- Oubeidillah, A. A., S.-C. Kao, M. Ashfaq, B. S. Naz, and G. Tootle, 2014: A large-scale, high-resolution hydrological model parameter data sets for climate change impact assessment for the conterminous US. *Hydrol. Earth Syst. Sci.*, **18**, 67–84, <https://doi.org/10.5194/hess-18-67-2014>.
- Pan, M., and Coauthors, 2003: Snow process modeling in the North American Land Data Assimilation System (NLDAS): 2. Evaluation of model simulated snow water equivalent. *J. Geophys. Res.*, **108**, 8850, <https://doi.org/10.1029/2003JD003994>.
- , A. K. Sahoo, T. J. Troy, R. K. Vinukollu, J. Sheffield, and E. F. Wood, 2012: Multisource estimation of long-term terrestrial water budget for major global river basins. *J. Climate*, **25**, 3191–3206, <https://doi.org/10.1175/JCLI-D-11-00300.1>.
- Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, 1992: *Numerical Recipes in FORTRAN 77: The Art of Scientific Computing*. 2nd ed. Cambridge University Press, 933 pp.
- Quiring, S., T. Ford, J. Wang, A. Khong, E. Harris, T. Lindgren, D. Goldberg, and Z. Li, 2016: The North American Soil Moisture Database: Development and applications. *Bull. Amer. Meteor. Soc.*, **97**, 1441–1459, <https://doi.org/10.1175/BAMS-D-13-00263.1>.
- Robock, A., and Coauthors, 2003: Evaluation of the North American Land Data Assimilation System over the southern Great Plains during the warm season. *J. Geophys. Res.*, **108**, 8846, <https://doi.org/10.1029/2002JD003245>.
- , K. Y. Vinnikov, G. Srinivasan, J. K. Entin, S. E. Hollinger, N. A. Speranskaya, S. Liu, and A. Namkhai, 2000: The Global Soil Moisture Data Bank. *Bull. Amer. Meteor. Soc.*, **81**, 1281–1299, [https://doi.org/10.1175/1520-0477\(2000\)081<1281:TGSMDB>2.3.CO;2](https://doi.org/10.1175/1520-0477(2000)081<1281:TGSMDB>2.3.CO;2).
- Royer A., and S. Poirier, 2010: Surface temperature spatial and temporal variations in North America from homogenized satellite SMMR-SSM/I microwave measurements and re-analysis for 1979–2008. *J. Geophys. Res.*, **115**, D08110, <https://doi.org/10.1029/2009JD0127>.
- Sahoo, A. K., M. Pan, T. J. Troy, R. Vinukollu, J. Sheffield, and E. F. Wood, 2011: Reconciling the global terrestrial water budget using satellite remote sensing. *Remote Sens. Environ.*, **115**, 1850–1865, <https://doi.org/10.1016/j.rse.2011.03.009>.
- Schaefer, G., M. Cosh, and T. Jackson, 2007: The USDA natural resources conservation service Soil Climate Analysis Network (SCAN). *J. Atmos. Oceanic Technol.*, **24**, 2073–2077, <https://doi.org/10.1175/2007JTECHA930.1>.
- Schroeder, J. L., W. S. Burgett, K. B. Haynie, I. Sonmez, G. D. Skwira, A. L. Doggett, and J. W. Lipe, 2005: The West Texas Mesonet: A technical overview. *J. Atmos. Oceanic Technol.*, **22**, 211–222, <https://doi.org/10.1175/JTECH-1690.1>.
- Scott, B. L., T. E. Ochsner, B. G. Illston, C. A. Fiebrich, J. B. Basara, and A. J. Sutherland, 2013: New soil property database improves Oklahoma Mesonet soil moisture estimates. *J. Atmos. Oceanic Technol.*, **30**, 2585–2595, <https://doi.org/10.1175/JTECH-D-13-00084.1>.
- Sheffield, J., and E. F. Wood, 2007: Characteristics of global and regional drought, 1950–2000: Analysis of soil moisture data from off-line simulation of the terrestrial hydrologic cycle. *J. Geophys. Res.*, **112**, D17115, <https://doi.org/10.1029/2006JD008288>.
- , C. R. Ferguson, T. J. Troy, E. F. Wood, and M. G. McCabe, 2009: Closing the terrestrial water budget from satellite remote sensing. *Geophys. Res. Lett.*, **36**, L07403, <https://doi.org/10.1029/2009GL037338>.
- Silver, N. C., J. Ullman, and C. J. Picker, 2015: COMPCOR: A computer program for comparing correlations using confidence intervals. *Psychol. Cognit. Sci. Open J.*, **1**, 26–28, <https://doi.org/10.17140/PCSOJ-1-104>.
- Su, Z., 2002: The Surface Energy Balance System (SEBS) for estimation of turbulent heat fluxes. *Hydrol. Earth Syst. Sci.*, **6**, 85–99, <https://doi.org/10.5194/hess-6-85-2002>.
- Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.*, **106**, 7183–7196, <https://doi.org/10.1029/2000JD900719>.
- Troy, T. J., E. F. Wood, and J. Sheffield, 2008: An efficient calibration method for continental-scale land surface modeling. *Water Resour. Res.*, **44**, W09411, <https://doi.org/10.1029/2007WR006513>.
- Velpuri, N. M., G. B. Senay, R. K. Singh, S. Bohms, and J. P. Verdin, 2013: A comprehensive evaluation of two MODIS evapotranspiration products over the conterminous United States: Using point and gridded FLUXNET and water balance ET. *Remote Sens. Environ.*, **139**, 35–49, <https://doi.org/10.1016/j.rse.2013.07.013>.
- Wood, E. F., D. P. Lettenmaier, X. Liang, B. Nijssen, and S. W. Wetzel, 1997: Hydrological modeling of continental-scale basins. *Annu. Rev. Earth Planet. Sci.*, **25**, 279–300, <https://doi.org/10.1146/annurev.earth.25.1.279>.
- , and Coauthors, 1998: The Project for Intercomparison of Land-Surface Parameterization Scheme (PILPS) Phase 2(c) Red–Arkansas River basin experiment: 1. Experiment description and summary intercomparisons. *Global Planet. Change*, **19**, 115–136, [https://doi.org/10.1016/S0921-8181\(98\)00044-7](https://doi.org/10.1016/S0921-8181(98)00044-7).
- Xia, Y., A. J. Pitman, H. V. Gupta, M. Leplastrier, A. Henderson-Sellers, and L. Bastidas, 2002: Calibrating a land surface model of varying complexity using multicriteria methods and the Cabauw dataset. *J. Hydrometeorol.*, **3**, 181–194, [https://doi.org/10.1175/1525-7541\(2002\)003<0181:CALSMO>2.0.CO;2](https://doi.org/10.1175/1525-7541(2002)003<0181:CALSMO>2.0.CO;2).
- , and Coauthors, 2012a: 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. *J. Geophys. Res.*, **117**, D03109, <https://doi.org/10.1029/2011JD016048>.

- , and Coauthors, 2012b: 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. *J. Geophys. Res.*, **117**, D03110, <https://doi.org/10.1029/2011JD016051>.
- , and Coauthors, 2013: Validation of Noah-simulated soil temperature in the North American Land Data Assimilation System phase 2. *J. Appl. Meteor. Climatol.*, **52**, 455–471, <https://doi.org/10.1175/JAMC-D-12-033.1>.
- , J. Sheffield, M. B. Ek, J. Dong, N. Chaney, H. Wei, J. Meng, and E. F. Wood, 2014: Evaluation of multi-model simulated soil moisture in NLDAS-2. *J. Hydrol.*, **512**, 107–125, <https://doi.org/10.1016/j.jhydrol.2014.02.027>.
- , M. B. Ek, Y. Wu, T. Ford, and S. M. Quiring, 2015a: Comparison of NLDAS-2 simulated and NASMD observed daily soil moisture. Part I: Comparison and analysis. *J. Hydrometeorol.*, **16**, 1962–1980, <https://doi.org/10.1175/JHM-D-14-0096.1>.
- , M. T. Hobbins, Q. Mu, and M. B. Ek, 2015b: Evaluation of NLDAS-2 evapotranspiration against tower flux site observations. *Hydrol. Processes*, **29**, 1757–1771, <https://doi.org/10.1002/hyp.10299>.
- , T. W. Ford, Y. Wu, S. M. Quiring, and M. B. Ek, 2015c: Automated quality control of in situ soil moisture from the North American Soil Moisture Database using NLDAS-2 products. *J. Appl. Meteor. Climatol.*, **54**, 1267–1282, <https://doi.org/10.1175/JAMC-D-14-0275.1>.
- , and Coauthors, 2016a: Basin-scale assessment of the land surface water budget in the National Centers for Environmental Prediction operational and research NLDAS-2 systems. *J. Geophys. Res. Atmos.*, **121**, 2750–2779, <https://doi.org/10.1002/2015JD023733>.
- , B. A. Cosgrove, K. E. Mitchell, C. D. Peters-Lidard, M. B. Ek, S. Kumar, D. Mocko, and H. Wei, 2016b: Basin-scale assessment of the land surface energy budget in the National Centers for Environmental Prediction operational and research NLDAS-2 systems. *J. Geophys. Res. Atmos.*, **121**, 196–220, <https://doi.org/10.1002/2015JD023889>.
- , D. M. Mocko, M. Huang, B. Li, M. Rodell, K. E. Mitchell, X. Cai, and M. Ek, 2017: Comparison and assessment of three advanced land surface models in simulating terrestrial water storage components over the United States. *J. Hydrometeorol.*, **18**, 625–649, <https://doi.org/10.1175/JHM-D-16-0112.1>.
- Yapo, P. O., H. V. Gupta, and S. Sorooshian, 1998: Multi-objective global optimization for hydrologic models. *J. Hydrol.*, **204**, 83–97, [https://doi.org/10.1016/S0022-1694\(97\)00107-8](https://doi.org/10.1016/S0022-1694(97)00107-8).
- Zhang, T., P. W. Stackhouse Jr., S. K. Gupta, S. J. Cox, J. C. Mikovitz, and L. M. Hinkelman, 2013: The validation of the GEWEX SRB surface shortwave flux data products using BSRN measurements: A systematic quality control, production and application approach. *J. Quant. Spectrosc. Radiat. Transfer*, **122**, 127–140, <https://doi.org/10.1016/j.jqsrt.2012.10.004>.
- , —, J. S. Gupta, S. J. Cox, and J. C. Mikovitz, 2015: The validation of the GEWEX SRB surface longwave flux data products using BSRN measurements. *J. Quant. Spectrosc. Radiat. Transfer*, **150**, 134–147, <https://doi.org/10.1016/j.jqsrt.2014.07.013>.
- Zhang, Y., and Coauthors, 2018: A Climate Data Record (CDR) for the global terrestrial water budget: 1984–2010. *Hydrol. Earth Syst. Sci.*, **22**, 241–263, <https://doi.org/10.5194/hess-22-241-2018>.
- Zou, G. Y., 2007: Toward using confidence intervals to compare correlations. *Psychol. Methods*, **12**, 399–413, <https://doi.org/10.1037/1082-989X.12.4.399>.