# Comparison and Assessment of Three Advanced Land Surface Models in Simulating Terrestrial Water Storage Components over the United States

YOULONG XIA,[a] DAVID MOCKO,[b] MAOYI HUANG,[c] BAILING LI,[d] MATTHEW RODELL,[b] KENNETH E. MITCHELL,[e] XITIAN CAI,[f] AND MICHAEL B. EK[g]

[a] *I. M. Systems Group at NOAA/NCEP/Environmental Modeling Center, College Park, Maryland*
[b] *Hydrological Science Laboratory, NASA Goddard Space Flight Center, Greenbelt, Maryland*
[c] *Pacific Northwest National Laboratory, Richland, Washington*
[d] *Earth System Science Interdisciplinary Center, University of Maryland, College Park, College Park, Maryland*
[e] *Prescient Weather Ltd., State College, Pennsylvania*
[f] *Department of Civil and Environmental Engineering, Princeton University, Princeton, New Jersey*
[g] *NOAA/NCEP/Environmental Modeling Center, College Park, Maryland*

(Manuscript received 13 May 2016, in final form 28 November 2016)

ABSTRACT

To prepare for the next-generation North American Land Data Assimilation System (NLDAS), three advanced land surface models [LSMs; i.e., Community Land Model, version 4.0 (CLM4.0); Noah LSM with multiphysics options (Noah-MP); and Catchment LSM-Fortuna 2.5 (CLSM-F2.5)] were run for the 1979–2014 period within the NLDAS-based framework. Unlike the LSMs currently executing in the operational NLDAS, these three advanced LSMs each include a groundwater component. In this study, the model simulations of monthly terrestrial water storage anomaly (TWSA) and its individual water storage components are evaluated against satellite-based and in situ observations, as well as against reference reanalysis products, at basinwide and statewide scales. The quality of these TWSA simulations will contribute to determining the suitability of these models for the next phase of the NLDAS. Overall, it is found that all three models are able to reasonably capture the monthly and interannual variability and magnitudes of TWSA. However, the relative contributions of the individual water storage components to TWSA are very dependent on the model and basin. A major contributor to the TWSA is the anomaly of total column soil moisture content for CLM4.0 and Noah-MP, while the groundwater storage anomaly is the major contributor for CLSM-F2.5. Other water storage components such as the anomaly of snow water equivalent also play a role in all three models. For each individual water storage component, the models are able to capture broad features such as monthly and interannual variability. However, there are large intermodel differences and quantitative uncertainties, which are motivating follow-on investigations in the NLDAS Science Testbed developed by the NASA and NCEP NLDAS teams.

## 1. Introduction

After more than 15 years of research and development for the North American Land Data Assimilation System (NLDAS) project (Mitchell et al. 2004), its second phase (NLDAS-2; Xia et al. 2012a,b) was implemented into the National Centers for Environmental Prediction (NCEP) operations in August 2014. This system is providing timely near-real-time operational products such as energy fluxes, water fluxes, and state variables to the user community, including various government agencies, academia, and private enterprises. Its major purpose is to support national operational drought monitoring and prediction tasks such as the U.S. Drought Monitor (Svoboda et al. 2002) and the Climate Prediction Center's Seasonal Drought Outlook. The current NLDAS-2 includes four land surface models (LSMs): the community Noah LSM (Ek et al. 2003), the Mosaic LSM (Koster and Suarez 1996), the Sacramento (SAC) Soil Moisture Accounting model (Burnash et al. 1973), and the Variable Infiltration Capacity model (VIC) (Liang et al. 1994). However, none of these LSMs includes a groundwater module, which is a very important component to represent the interaction between deep soil water and an unconfined aquifer (Niu et al. 2011; Barlage et al. 2015). In addition, the simple one- or two-layer snow models used for snowpack simulation in these LSMs lead to unrealistic runoff simulations,

*Corresponding author e-mail*: Youlong Xia, youlong.xia@noaa.gov

DOI: 10.1175/JHM-D-16-0112.1

especially the timing of the peak in monthly runoff in the spring over the mountainous western regions (Xia et al. 2012b; Cai et al. 2014b). These weaknesses are being addressed by the LSM development community by adding and enhancing model physical processes to the NLDAS.

The Community Land Model (CLM) was developed as the land component of the Community Climate System Model [now known as the Community Earth System Model (CESM)]. It has various versions (Bonan 1996; Oleson and Bonan 2000; Dickinson et al. 2006; Oleson et al. 2008, 2010). In this study, Climate Land Model, version 4.0 (CLM4.0; Lawrence et al. 2011), is used. Noah LSM with multiphysics options (Noah-MP; Niu et al. 2011; Yang et al. 2011) is developed by augmenting the original Noah model with multiple physical process enhancements is used in this study. The Noah-MP has been coupled to the Weather Research and Forecasting (WRF) Model (Barlage et al. 2015) and is used in both the uncoupled and coupled WRF-Hydro systems (Gochis et al. 2013; Senatore et al. 2015).

The Mosaic model has been upgraded to the Catchment LSM (CLSM) by adding a shallow groundwater module and enhancing its hydrology using a catchment-based approach (Koster et al. 2000). As a key component of NASA's Earth model system, CLSM has been widely used in both uncoupled reanalysis systems such as the Modern-Era Retrospective Analysis for Research and Applications (MERRA; Rienecker et al. 2011) and fully coupled global Earth system models such as the Goddard Earth Observing System, version 5 (GEOS-5; De Lannoy et al. 2013). The CLSM-Fortuna 2.5 (CLSM-F2.5) used in NASA GEOS-5 (De Lannoy and Reichle 2016) is used in this study.

Although CLM4.0 and Noah-MP have been executed in the NLDAS test bed over the past several years and their performance has been evaluated in several aspects (Cai et al. 2014b), their water storage components such as terrestrial water storage, snow water equivalent (SWE), total column soil moisture content, and groundwater storage have not yet been comprehensively evaluated thus far. For example, it has been recognized that a constant value of 0.2 for specific yield value (a factor to convert water-table depth to groundwater storage) to linearly convert water-table depth to groundwater storage for CLM4.0 and Noah-MP is not justified as it should spatially vary based on regional hydrogeological conditions (Cai et al. 2014b; Huang et al. 2013; Ren et al. 2016). Therefore, a reevaluation for terrestrial water storage and groundwater storage is needed by using the groundwater storage that is directly output from these two LSMs, as we do here. Moreover, CLSM-F2.5 has not yet been evaluated in the NLDAS test bed, although it is the advanced land surface model intended to replace the Mosaic model. To achieve the next-generation NLDAS, selection of more advanced land surface models is a critical task and is the first priority. Other priorities important to advancing the NLDAS are upgrading soil and vegetation parameters, improving the surface meteorological forcing, executing on a finer spatial resolution, and adding data assimilation capabilities.

Recently, a number of new reference datasets have become available for use in assessment of land models and Land Data Assimilation Systems globally and over the continental United States (CONUS). These provide new opportunities to comprehensively evaluate/reevaluate the performance of the three LSMs. These reference datasets include the terrestrial water storage anomaly (TWSA) from the Gravity Recovery and Climate Experiment (GRACE; Wahr et al. 1998; Landerer and Swenson 2012), in situ soil moisture observations collected from the North American Soil Moisture Database (NASMD; Xia et al. 2015c; Quiring et al. 2016), in situ well measurements from the U.S. Geological Survey (Rodell et al. 2007; Li et al. 2015), and SWE reanalysis data from the National Weather Service's Snow Data Assimilation System (SNODAS; Clow et al. 2012; Artan et al. 2013).

In line with the NLDAS-2 configuration (Xia et al. 2012a) and Land Information System (LIS) framework (Kumar et al. 2006), all three LSMs are driven by the same hourly NLDAS-2 surface meteorological forcing for the period 1979–2014 over the United States at 0.125° spatial resolution. Section 2 describes the three LSMs and their configurations, the experimental design, surface forcing data, model output, and the model evaluation criteria. Section 3 introduces the observed and reference datasets. Section 4 presents the specific evaluations for various water storage components. Section 5 discusses the relative contributions of different components to the TWSA, as well as possible applications. Section 6 gives final conclusions and suggestions for future studies.

## 2. Models and forcing data

### a. Models

The three LSMs used in this study are CLM4.0, Noah-MP, and CLSM-F2.5. An important common feature of these models is that they contain a groundwater module. Soil depth varies from 3.8 m in CLM4.0 and 2 m in Noah-MP to 1 m in CLSM-F2.5. The corresponding unconfined aquifer depth varies from model to model with the deepest being CLM4.0 (i.e., 5 m) and the shallowest being CLSM-F2.5 (i.e., 3–6 m; Fig. 1). Table 1 compares the attributes of these three models from the aspect of surface energy and water cycles. Additional model technical details are described below.

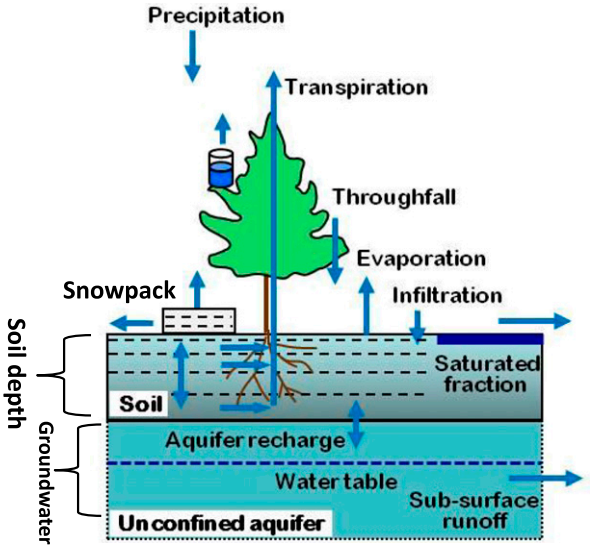


FIG. 1. Schematic diagram describing model structure and water cycling processes for three LSMs [modified from Lawrence et al. (2011)]. The number of soil layers and soil depth vary among the three models (see Table 1). Note that CLM4.0 and Noah-MP physics are executed for uniform grid cells, while CLSM-F2.5 is executed for hydrological catchment units.

### 1) CLM4.0

CLM has a successful development history from its original version (Bonan 1996) to its version 4.0 (Lawrence et al. 2011) as a result of a community development effort. Compared to its previous versions, CLM4.0 was upgraded by augmenting the representations of various hydrological processes, including those associated with soil hydrology, runoff generation, groundwater dynamics, and snowpack. The 3.8-m soil depth is divided into 10 hydrologically active layers with varying soil layer thicknesses (see Table 1). Below the 3.8-m soil depth, the 38.2-m-thick ground is divided into five bottom thermal layers that are not hydrologically active to accurately capture soil temperature dynamics in century-scale integrations (Alexeev et al. 2007). Therefore, CLM4.0 has a total soil depth of 3.8 m for hydrological processes and 42 m for thermodynamic processes. Soil water is calculated by using a revised numerical solution of the one-dimensional Richards equation (Zeng and Decker 2009). Overland flow is computed by using a simplified TOPMODEL-based representation that considers subgrid variability of surface topography and its impact on runoff generation through the Dunne mechanism (Niu et al. 2005). Base flow is calculated as an exponential function of the grid-mean water-table depth as used in the original TOPMODEL, in which the water-table depth is inferred from the aquifer water storage scaled by the average specific yield as described in Niu et al. (2007). The aquifer is assumed to have a storage capacity of 5000 mm. Because of the lack of data to constrain the specific yield, it is set to be 0.2 globally (Oleson et al. 2010). The soil hydrology scheme of CLM4.0 includes the exchange of water between an unconfined aquifer and the overlying soil column (Niu et al. 2007). The five-layer snow model is updated with new snow cover and snow burial fraction calculated by using the

TABLE 1. Primary attributes of the three LSMs used in this study.

| Description | CLM4.0 (Lawrence et al. 2011) | Noah-MP (Niu et al. 2011) | CLSM-F2.5 (Koster et al. 2000) |
|---|---|---|---|
| Energy balance | Yes | Yes | Yes |
| Water balance | Yes | Yes | Yes |
| Model time step | 1 h | 1 h | 1 h |
| No. of model soil layers | 10 | 4 | 2 |
| Depth of total soil column (m) | 3.8 | 2.0 | 1.0 |
| Model soil layer thickness (m) | 0.018, 0.028, 0.045, 0.075, 0.124, 0.204, 0.336, 0.553, 0.913, 1.506 | 0.1, 0.3, 0.6, 1.0 | 0.02, 1.0 |
| Tiling: vegetation | Yes | Yes | Yes |
| No. of snow model layers | 5 | 3 | 3 |
| Frozen soil thermal | Yes | Yes | Yes |
| Frozen soil hydraulics | Yes | Yes | Yes |
| Soil thermodynamics | Heat conduction equation | Heat conduction equation | Heat conduction equation |
| Soil temperature profile | Yes | Yes | Yes |
| Soil water drainage | Yes | Yes | Yes |
| Soil water vertical diffusion | Yes | Yes | Yes |
| TOPMODEL for surface runoff | Yes | Yes | Yes |
| No. of canopy layers | 1 | 1 | 0 |
| Dynamical vegetation | Yes (Dickinson et al. 1998) | Yes (Dickinson et al. 1998) | No |
| Explicit vegetation | Yes | Yes | Yes |
| Canopy resistance | Ball–Berry algorithm (Ball et al. 1987) | Ball–Berry algorithm (Ball et al. 1987) | Jarvis algorithm (Jarvis 1976) |

parameterization schemes of Niu and Yang (2006) and Wang and Zeng (2009). The Snow, Ice, and Aerosol Radiative (SNICAR) model describing grain size–dependent snow aging and vertically resolved snowpack heating is calculated following Flanner et al. (2007).

CLM4.0 contains spatial heterogeneity of surface land-cover types, including glacier, lake, wetland, urban, and vegetated surfaces. The latter is further decomposed into 15 plant functional types (PFTs) and bare ground. In this study, the percentages of surface land-cover types and PFTs within a 0.125° grid cell are derived from the 0.05° Moderate Resolution Imaging Spectroradiometer (MODIS)-based global land parameter dataset in 2005 (Ke et al. 2012). Soil texture is produced from a hybrid of the 30-arc-s State Soil Geographic Database (STATSGO; Miller and White 1998). The two-layer soil type data are converted to a composition of clay and sand (Cosby et al. 1984) within each 30-arc-s grid cell to 10 vertical layers down to a depth of 3.8 m.

#### 2) NOAH-MP

Several recent LSM physics improvements were added to the early Noah model (Ek et al. 2003) to construct the newer Noah-MP (Niu et al. 2011; Yang et al. 2011). The major Noah LSM augmentations include the following: 1) modifying the model structure to include a one-layer canopy and three-layer snowpack (the Noah LSM has no explicit canopy layer and includes only a one-layer bulk snowpack), 2) adding a tiling scheme (using different tiles for vegetation and bare soil) that separates vegetated areas from bare ground to better calculate the land surface energy balance (the Noah LSM applies a single vegetation type for each grid cell, but also uses the green vegetation fraction to weight the contribution of evapotranspiration from vegetation and direct evaporation from bare ground to the total evaporation), 3) incorporating a more reasonable permeable frozen soil scheme by separating permeable and impermeable fractions (Yang and Niu 2003), 4) adding a TOPMODEL-based runoff scheme (Niu et al. 2005) and a Simple Groundwater Model (SIMGM; Niu et al. 2007) to improve the modeling of soil hydrology, and 5) adding a short-term leaf dynamic model modified from Dickinson et al. (1998) to simulate leaf area index (LAI). Noah-MP contains multiparameterization options for dynamic vegetation, canopy stomatal resistance, runoff generation, a groundwater module, and other physical processes. Based on previous experiments from Niu and Yang (2006), the schemes used in this study are dynamic vegetation, the Ball–Berry scheme for canopy stomatal resistance, TOPMODEL runoff calculation with SIMGM, the Monin–Obukhov scheme for the surface exchange coefficient for heat and moisture, the frozen soil scheme from Niu and Yang (2006), and a modified two-stream scheme for the transfer of radiation through the vegetation canopy. Noah-MP has been evaluated at regional (Niu et al. 2011; Cai et al. 2014a; 2014b) and global scales (Yang et al. 2011).

Noah-MP applies the Canadian Land Surface Scheme type for snow cover albedo (Verseghy 1991), the STATSGO soil texture data, and the 0.144° global monthly snow-free albedo derived by Csiszar and Gutman (1999). Last, Noah-MP keeps the same four soil layers and layer thicknesses (see Table 1) as in the previous Noah version.

#### 3) CLSM-F2.5

CLSM-F2.5 is a land model intended for the land component of global-scale coupled land–atmosphere modeling (Koster et al. 2000). Unlike CLM4.0 and Noah-MP, the model's basic computational unit is the hydrological catchment/watershed rather than a uniform latitude–longitude grid cell. In each catchment the vertical profile of soil moisture is determined by an equilibrium soil moisture profile from the surface to the water table and by two additional variables that describe deviations from the equilibrium profile in a 1-m root-zone layer and a 2-cm surface layer, respectively. To increase the subgrid heterogeneity of soil moisture, each catchment is separated into three distinct and dynamically varying subareas: a saturated region, an unsaturated region, and a wilting region where different runoff and evapotranspiration schemes are applied. It must be noted that CLSM-F2.5 does not explicitly simulate the water table. The catchment deficit variable (amount of water needed to fill a catchment) reflects water changes in the entire catchment including shallow unconfined aquifer. Groundwater storage can be derived from the catchment deficit and the maximum water capacity of each catchment, which is determined by the bedrock depth parameter and soil porosity. Following Houborg et al. (2012), the bedrock depths in the CLSM-F2.5 here were increased uniformly by 2 m everywhere so that the dynamic range of simulated TWSA would better agree with GRACE-derived TWSA, especially in dry periods (Li et al. 2012).

CLSM-F2.5 also includes a three-layer snowpack model that accounts for snow melting and refreezing, dynamic changes in snow density, snow insulating properties, and other physics relevant to the growth and ablation of the snowpack (Lynch-Stieglitz 1994). The snow model contains three prognostic variables (heat content, SWE, and snow depth) for each of three vertically stacked layers. The treatment of snow albedo uses different reflectance values for the visible (VIS) and near-infrared (NIR) radiation bands, with reductions in albedo imposed by both

vegetation masking and fractional snow cover (Hansen et al. 1983; Stieglitz et al. 2001). CLSM-F2.5 uses a modified Simple Biosphere Model (SiB; Sellers et al. 1996) albedo parameterization scheme to compute surface 1) visible direct, 2) near-infrared direct, 3) visible diffuse, and 4) infrared diffuse albedo at each model time step. The modified SiB scheme takes advantage of available MODIS albedo data to match the annual cycles of visible diffuse and infrared diffuse albedo of MODIS (Moody et al. 2008; Gao et al. 2014). Last, the CLSM-F2.5 runs for this study use 1) the National Geophysical Data Center (NGDC) soil texture data (Reynolds et al. 2000) with soil hydraulic parameters from the Second Global Soil Wetness Project (GSWP-2; Dirmeyer et al. 2006), 2) the vegetation classification data derived from University of Maryland (UMD) Advanced Very High Resolution Radiometer (AVHRR) on NOAA polar satellites global 1-km vegetation database, and 3) AVHRR monthly GSWP-2 (Dirmeyer et al. 2006) LAI and greenness fraction climatology.

### b. *Surface meteorological forcing data*

The 36-yr (1979–2014) hourly land surface forcing data produced for NLDAS-2 (Xia et al. 2012a) are used to drive the three land models. This surface forcing data include gauge-based precipitation, 2-m air temperature and specific humidity, surface pressure, downward shortwave and longwave radiation, and 10-m wind speed. The primary data source is the 36-yr, 3-hourly, 32-km output of North American Regional Reanalysis (NARR; Mesinger et al. 2006). All of these NARR outputs are temporally interpolated to hourly resolution and spatially interpolated to a 0.125° NLDAS grid. To account for impact of elevation differences between the NLDAS grid and the NARR grid for air temperature, humidity, and downward longwave radiation, a terrain height adjustment is applied to the air temperature using a uniform lapse rate of $6.5\,\mathrm{K\,km^{-1}}$. The specific humidity and downward longwave radiation are then adjusted by the adjusted air temperature (Cosgrove et al. 2003a).

The NLDAS-2 precipitation is anchored to the Climate Prediction Center (CPC) 0.125°, gauge-only, daily U.S. precipitation analysis with bias correction from the monthly Parameter-Elevation Regressions on Independent Slopes Model (PRISM; Daly et al. 1994). The daily NLDAS-2 precipitation is temporally disaggregated to hourly values by using hourly temporal weights derived from either the NOAA/NCEP/Environmental Modeling Center hourly stage II radar radiation (Lin and Mitchell 2005), CPC morphing technique (CMORPH) hourly precipitation (Joyce et al. 2004), CPC coarse-resolution hourly precipitation, or the interpolated hourly NARR precipitation in that order based on data availability. The temporally interpolated downward shortwave radiation is bias corrected by a ratio-based method (Berg et al. 2003) that applies the UMD's Surface Radiation Budget (SRB) dataset (Pinker et al. 2003).

### c. *Model spinup procedure*

For LSM simulations in general, the initial conditions (e.g., soil moisture) have a significant impact on water fluxes, energy fluxes, and state variables. To minimize this impact, an individual LSM needs to run many years (typically called a spinup run) to achieve its own equilibrium state. Yang et al. (1995) investigated the spinup time required for 22 LSMs and found that the models took about 12 years to reach the equilibrium for tropical forest and grassland sites. Cosgrove et al. (2003b) discussed the LSM spinup behavior in NLDAS and found that the models need 2–10 years to reach their own equilibrium states depending on the model and climate region (dry or wet region). It should be noted that the models discussed above do not include a groundwater module. For those LSMs with groundwater dynamics, the required spinup time is over 70 years (Niu et al. 2007; Cai et al. 2014b).

In this study, CLM4.0 was retrospectively run by recycling the NLDAS-2 forcing from 1979 to 2007 for 36 cycles (~1000 years) until all state variables including 10-layer soil moisture, 15-layer soil temperature, and water-table depth reached their equilibrium. This very long spinup period is needed for CLM4.0 because it has a very deep soil column and water table, which requires a substantially long time period for soil temperature and moisture to reach equilibrium, especially in cold and dry regions (Liang et al. 2003). The Noah-MP and CLSM-F2.5 were initialized by running the models repeatedly two times through a 35-yr period (1979–2013). For both Noah-MP and CLSM-F2.5 simulations, nearly 100% of all NLDAS grid points achieve an equilibrium state as the change of terrestrial water storage values is within 1% after 70 years of spinup.

### d. *Model output*

After the model spinup runs were completed, the LSMs were executed from 1 January 1979 to 31 December 2014 using NLDAS-2 hourly meteorological surface forcing. The 36-yr monthly total column soil moisture content, top 1-m soil moisture, SWE, and groundwater storage change were output from all three LSMs.

The TWSA is calculated as the sum of 1) total column soil moisture content anomaly (SMCA), 2) snow water equivalent anomaly (SWEA), and 3) groundwater storage anomaly (GWSA). We use a common inland mask to exclude all grid cells of surface water bodies such as rivers, lakes, ponds, and reservoirs for both GRACE and the

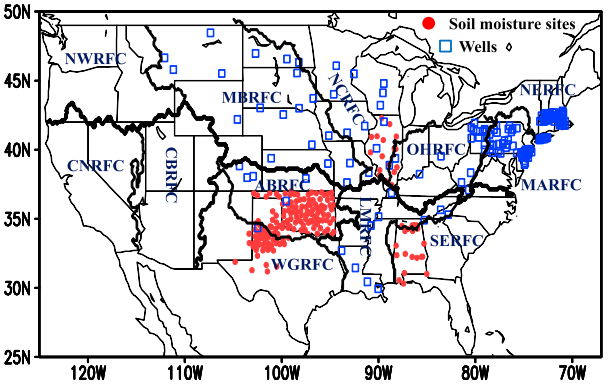


FIG. 2. Names and boundaries of the domains of 12 NWS RFCs, locations of 181 USGS wells, and locations of 195 in situ soil moisture observational sites (CBRFC, Colorado; CNRFC, California–Nevada; WGRFC, West Gulf; MBRFC, Missouri; ABRFC, Arkansas; NCRFC, North-Central; NWRFC, Northwest; MARFC, Mid-Atlantic; SERFC, Southeast; NERFC, Northeast; LMRFC, Lower Mississippi; and OHRFC, Ohio).

models as there is no explicit realistic representation of those water bodies in the three models of this study. This exclusion is not realistic; however, it is reasonable as these water bodies are excluded from the calculation of water balance in most LSMs, such as the LSMs in this study.

### e. Evaluation strategy

The TWSA and its component anomalies (i.e., SMCA, SWEA, and GWSA) are evaluated against GRACE-based TWSA, U.S. Geological Survey (USGS)-measured GWSA, and SNODAS-reanalyzed SWEA over 6–12 National Weather Service (NWS) River Forecast Centers (RFCs; Fig. 2), depending on which water storage component is assessed. RFC basins are selected here as the assessment spatial scale because their sizes are sufficient to meet the requirement for application of GRACE TWSA data. GRACE data are useful for basins with a typical basin size larger than ~1 000 000 km$^2$ (Bingham and Hughes 2006; Syed et al. 2005; Wahr et al. 2004, 2006), although GRACE data were used recently for some smaller basins with an area size between 200 000 and 1 000 000 km$^2$ (Long et al. 2014; Strassberg et al. 2009; Swenson et al. 2006, 2008; Xiao et al. 2015). Moreover, RFC basins are an appealing spatial averaging choice, because the hydroclimatology across a given RFC basin is reasonably uniform spatially. As soil moisture measurements for the top 1 m are very sparse over both temporal and spatial scales (Xia et al. 2015c; Quiring et al. 2016), only monthly data from four states/regions (i.e., Alabama, Illinois, Oklahoma, and western Texas) are used in this study.

To assess model simulations of monthly TWSA and its individual water storage components, we utilize the mean square error (denoted as $E^2$) of the time series of the monthly anomaly with respect to the observed anomaly. We follow the common practice of expressing $E^2$ as the sum of 1) the square of the standard deviation of the error and 2) the square of the bias error:

$$E^2 = E_p^2 + E_m^2, \tag{1a}$$

$$E_p^2 = \frac{\sum_{i=1}^{N} (S_i' - O_i')^2}{N}, \tag{1b}$$

$$\gamma = \frac{\sigma_S}{\sigma_O}, \tag{1c}$$

$$E_m^2 = (\overline{S} - \overline{O})^2, \tag{1d}$$

$$\sigma_S = \sqrt{\frac{1}{N-1} \sum_{i=1}^{i=N} (S_i - \overline{S})^2} \quad \text{and}$$

$$\sigma_O = \sqrt{\frac{1}{N-1} \sum_{i=1}^{i=N} (O_i - \overline{O})^2}, \quad \text{and} \tag{1e}$$

$$S_i' = S_i - \overline{S} \quad \text{and} \quad O_i' = O_i - \overline{O}, \tag{1f}$$

where $E_m$ is the traditional bias error and $\sigma_S$ and $\sigma_O$ are the standard deviation (not error standard deviation) of the simulated data (i.e., model) and observed data, respectively. Variables $S_i$ and $O_i$ ($i = 1, N$) are time series of monthly model data and observation data, respectively. The parameter $\gamma$ is the ratio between the modeled and observed standard deviation. Variables $\overline{S}$ and $\overline{O}$ are multiyear mean modeled and observed data and $S_i'$ and $O_i'$ are time series of model and observation monthly anomalies (with respect to their multiyear mean). Variable $N$ is total number of months. Variable $E_p$ is used to measure the errors between modeled and observed data anomaly. Additionally, we use $\gamma$ to evaluate how well the observed variance is modeled. When $\gamma < 1$, this indicates the simulated anomaly is smoother than the observed anomaly, while $\gamma > 1$ indicates the simulated anomaly has larger variability than the observed anomaly. The bias portion $E_m$ measures how far the mean of modeled data is from the mean of observed data. Variable $E_m$ will not be discussed in this study because 1) GRACE provides only TWSA data and 2) the mean bias of the model simulation can be easily removed by a bias-correction procedure. Hence, in this setting, a perfect simulation means that $E_p = 0$ and $\gamma = 1$.

In this study, two anomaly time series are used for analysis and comparison. The first anomaly time series are calculated as the differences between the raw time series and long-term mean values as presented in Eq. (1f). They are used for analyses and calculations, except anomaly correlation (AC) calculation. The second anomaly time series are computed as the differences between the raw

TABLE 2. RFC name, data period, number of wells, average specific yield $\overline{S}_y$, average well depth $\overline{d}_{\text{well}}$, average depth to groundwater $\overline{d}_{\text{gw}}$, and average annual NLDAS precipitation $\overline{P}$.

| RFC name | Data period | No. of wells | $\overline{S}_y$ (unitless) | $\overline{d}_{\text{well}}$ (m) | $\overline{d}_{\text{gw}}$ (m) | $\overline{P}$ (mm) |
|---|---|---|---|---|---|---|
| MBRFC | 1980–2010 | 19 | 0.14 | 30 | 9 | 600 |
| AB-LMRFC | 1980–2010 | 13 | 0.16 | 86 | 17 | 970 |
| NCRFC | 1980–2010 | 13 | 0.17 | 19 | 6 | 900 |
| MARFC | 2002–11 | 62 | 0.12 | 35 | 8 | 1200 |
| NERFC | 1992–2011 | 64 | 0.23 | 12 | 6 | 1210 |
| OHRFC | 1980–2010 | 10 | 0.09 | 38 | 7 | 1190 |

time series and long-term mean seasonal cycles calculated from multiyear monthly mean time series. They are mainly used to calculate AC values that can reasonably represent the simulation skill and interannual variability.

## 3. Observed and reference datasets

### a. GRACE satellite–based TWSA

The twin satellites of GRACE measure the monthly changes in Earth's gravity field over space and time by using the variation of the distances between the two satellites (Tapley et al. 2004). Changes in the gravity field can be used to derive changes in (vertically integrated) terrestrial water storage. The effective resolution of GRACE data is around a few hundred kilometers because of the postprocessing techniques applied for increasing the signal to noise ratio (Rowlands et al. 2005; Swenson et al. 2006; Landerer and Swenson 2012).

In this study, we use the latest product release (version 05) of the GRACE-derived TWSA (RL05; Landerer and Swenson 2012). The data are obtained from the GRACE Tellus website (ftp://podaac-ftp.jpl.nasa.gov/allData/tellus/L3/land_mass/RL05/netcdf/). The RL05 version includes three products processed at the following three centers: the Center for Space Research (CSR) in the United States, the Jet Propulsion Laboratory (JPL) in the United States, and the German Research Centre for Geosciences (GFZ) in Germany. The anomalies of each of the three TWSA products from the Tellus website are calculated relative to each product's own long-term mean for the 2004–09 time period. The products feature a spatial discretization of 1° and a monthly time step. The RL05 product, spanning from December 2002 to January 2015, contains a total of 146 months of data with 10 months of missing data (i.e., June 2003; January and June 2011; May and October 2012; March, August, and September 2013; and February and December 2014). In this study we use a simple average of the CSR, GFZ, and JPL TWSA products as our observed TWSA value, following Sakumura et al. (2014), to reduce the noise in the three different products.

### b. USGS wells observations

The 181 USGS wells are located in the Missouri Basin RFC (MBRFC), the combined Arkansas Basin RFC and Lower Mississippi RFC (AB-LMRFC), the North-Central RFC (NCRFC), the Mid-Atlantic RFC (MARFC) covering only Pennsylvania and New Jersey, the Northeast RFC (NERFC) spanning only Massachusetts and New York, and the Ohio RFC (OBRFC; see Fig. 2 and Table 2). The observed well data have been used for validating GRACE-derived or model-estimated GWSAs (Rodell et al. 2007; Cai et al. 2014a; Li and Rodell 2015; Li et al. 2015). These wells were selected for exhibiting seasonal variation and minimal impact of pumping or injections. We obtained the monthly observations of water-table depth from the USGS Groundwater Watch website (http://groundwaterwatch.usgs.gov/), USGS National Water Information System, and the Illinois State Water Survey. The lengths of the data records range from 10 to over 30 years (Table 2). Monthly GWSAs were derived from water-table depth and specific yield values (see Li and Rodell 2015). The specific yield values $S_y$ were determined individually for each well based on published studies on the aquifer formation or, as a last resort, published $S_y$ estimates for the aquifer type. When multiple possible $S_y$ values were found for a given well, an $S_y$ within that range was selected based on the well depth and comparison of the dynamic range of water depths with those of neighboring wells. It should be pointed out that these $S_y$ values are the best estimates, not obtained through aquifer testing.

### c. SNODAS snow water equivalent

The SNODAS program provides daily gridded estimates of SWE, and related snow parameters, at a 1-km resolution for the United States (Clow et al. 2012). SNODAS combines data from various sources such as ground observations and airborne and satellites estimates with model results to produce a 1-km spatially distributed estimate of SWE (Carroll et al. 2006). Forcing data come from the NOAA Rapid Update Cycle 2 (RUC2) numerical weather prediction model

TABLE 3. Information on measurement networks, number of sites, data covering periods, and references of networks is given (USCRN is U.S. Climate Reference Network).

| Region name | No. of sites | Integrated soil depth (m) | Data covering period | Networks and reference |
|---|---|---|---|---|
| Alabama | 22 | 1 | From 1 Jan 2002 to 31 Dec 2012 (daily) | SCAN (Schaefer et al. 2007) and USCRN (Bell et al. 2013) |
| Illinois | 17 | 1, 2 | From Jan 1984 to Dec 2004 (monthly) | Illinois soil moisture databank (Robock et al. 2000) |
| Oklahoma | 97 | 1 | From 1 Jan 2000 to 31 Dec 2012 (daily) | Oklahoma Mesonet Network (Scott et al. 2013) |
| Western Texas | 59 | 1 | From 1 Jan 2000 to 31 Dec 2010 (daily) | West Texas Mesonet Network (Schroeder et al. 2005) |

output and are downscaled from 13-km into 1-km resolution using a digital elevation model. The multilayer snow model is an energy and mass balance model based on the Snow Thermal Model released in 1989 (SNTHERM.89; Jordan 1991). Assimilated observations include state and federal automated ground observations, snow surveys, and gamma flights as well as satellite-based snow extent information. As a publicly available, large-scale SWE product, it has been evaluated in many previous studies (Azar et al. 2008; Clow et al. 2012; Artan et al. 2013).These results show that the SNODAS is a good SWE reference dataset when compared with in situ observations. It is also used to evaluate LSM SWE simulations and satellite retrievals for SWE (Barlage et al. 2010; Vuyovich et al. 2014). We obtained 11-yr (2004–14) daily 1-km SNODAS SWE data from the National Snow and Ice Data Center website (http://nsidc.org/data). The monthly SWE is calculated from daily data and then monthly values are regridded to the 0.125° NLDAS grid.

### d. NASMD soil moisture content observations

NASMD (Quiring et al. 2016) is a collection of daily soil moisture values measured at about 1800 sites over 30 networks across the United States and Canada (Xia et al. 2015c). In situ soil moisture content has been quality controlled by excluding instrument malfunction errors, partially frozen soil cases, and other control processes. The details are further presented in Xia et al. (2015c) and Quiring et al. (2016). The daily data can be obtained from the NASMD website (http://soilmoisture.tamu.edu/Data/Download.aspx). Based on the length, quality, and measurement depth of data records, we selected four states/regions for our assessment: Alabama, Oklahoma, Illinois, and western Texas (see Fig. 2 and Table 3). The monthly soil moisture content for the top 1-m soil depth is calculated from the daily values based on the vertical integration of observations at multiple discrete depths. These data have been widely applied to many research studies (Xia et al. 2014, 2015a,b; Zhuo et al. 2015; Otkin et al. 2016).

## 4. Evaluation of modeled water storage component anomalies

### a. Evaluation of modeled TWSA using GRACE

Terrestrial water storage from GRACE includes contributions from soil moisture, snow and ice, groundwater, surface water (e.g., lakes and rivers), and vegetation water content. Its anomaly (TWSA) has been widely used to evaluate the performance of various LSMs (Güntner et al. 2007; Niu et al. 2007; Lo et al. 2010; Houborg et al. 2012; Leng et al. 2014; Cai et al. 2014a,b). Over the United States, the major term for terrestrial water storage (TWS) is soil moisture, SWE, groundwater, and surface water storage. As most LSMs (including the three LSMs studied herein) do not include lake and river modules, their terrestrial water storage is generally calculated as a sum of soil moisture, SWE, and groundwater storage when a groundwater module is included.

We used a simple mathematical average of the CSR, GFZ, and JPL products to represent the GRACE estimates. We calculated anomalies of modeled data relative to the same mean period as the Tellus website used for GRACE (i.e., 2004–09). We used an inland water mask, RFC mask, gridded modeled TWSA, and gridded GRACE TWSA to calculate the spatially averaged TWSA for the 12 RFCs (Fig. 2).

A comparison of the 12-yr (2003–14) mean seasonal cycle in Fig. 3 shows that the TWSA simulated from all three models is close to GRACE-observed TWSA for most of the 12 RFCs, suggesting reasonable model simulations. However, the CLSM-F2.5 simulation has relatively large differences from GRACE for MARFC and OHRFC, while the CLM4.0 simulation has relatively large differences for SERFC and OHRFC (see Fig. 2 for full RFC names). All three models depart

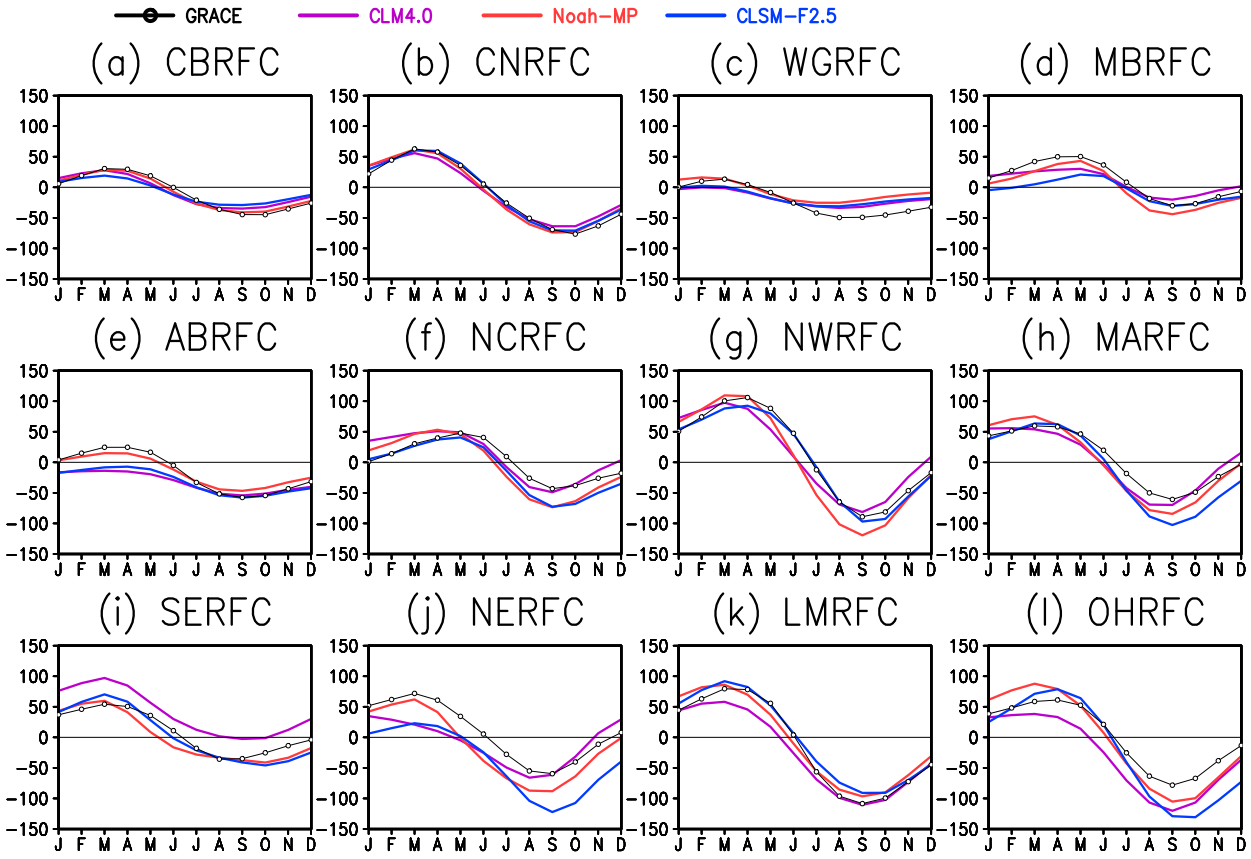


FIG. 3. Mean annual cycle of mean monthly TWSAs (mm) for each of 12 RFCs for three models and GRACE observations, calculated from a 12-yr (2003–14) monthly time series (thick black line with open circles, GRACE; purple line, CLM4.0; red line, Noah-MP; blue line, CLSM-F2.5).

notably from the GRACE seasonal cycle in NERFC. Clearly, each model has different degrees of performance depending on which RFC is examined. Thus, further effort is needed to improve the consistency of the modeled capability across the RFCs. Interestingly, in Fig. 3, the models seem better able to match the GRACE mean seasonal cycle in the drier RFCs, which have smaller amplitude of their seasonal cycle, than in the wetter RFCs.

In Fig. 4, the AC is typically larger than 0.5 (the NERFC is a notable exception with all AC $< 0.35$) and usually larger than 0.7 for all 12 RFCs, except for MARFC, OHRFC, and NERFC. It is evident that the simulation skill represented by AC depends on the model and basin. For example, CLM4.0 has the highest AC values in six RFCs (CBRFC, CNRFC, MBRFC, NCRFC, NWRFC, and MARFC), Noah-MP has the highest AC values in three RFCs (MARFC, NERFC, and OHRFC), and CLSM-F2.5 has the highest AC value in four RFCs (WGRFC, ABRFC, SERFC, and LMRFC; see Fig. 2 for full RFC names). However, each model's mean AC value averaged over the 12 RFCs is similar, specifically, 0.70 for Noah-MP, 0.69 for CLM4.0, and 0.68 for CLSM-F2.5 (Fig. 4). The reason why model performance varies substantially from model to model and basin to basin remains unclear and needs further investigation. However, the deeper soil column in CLM4.0 may help increase its AC values in relatively dry RFCs. All three models have better performance (larger AC values) for relatively dry RFCs (Figs. 4a–g) than for relatively wet RFCs (Figs. 4h,i).

We next examine in Fig. 5a the models' ability to simulate TWSA variability, as measured by the parameter $\gamma$ [Eq. (1c)]. All three models underestimate TWSA variability ($\gamma < 1$) in dry basins, while in wet basins the models tend to overestimate anomaly variability ($\gamma > 1$), especially in CLSM-F2.5, and less so in Noah-MP. Figure 6a, which presents the $E_p$ [Eq. (1b)] of simulated TWSA for each RFC with respect to GRACE data, illustrates that $E_p$ values are larger over the wet basins, with CLSM-F2.5 showing large errors for MARFC, NERFC, and OHRFC basins and CLM4.0 showing large error for OHRFC. Such larger

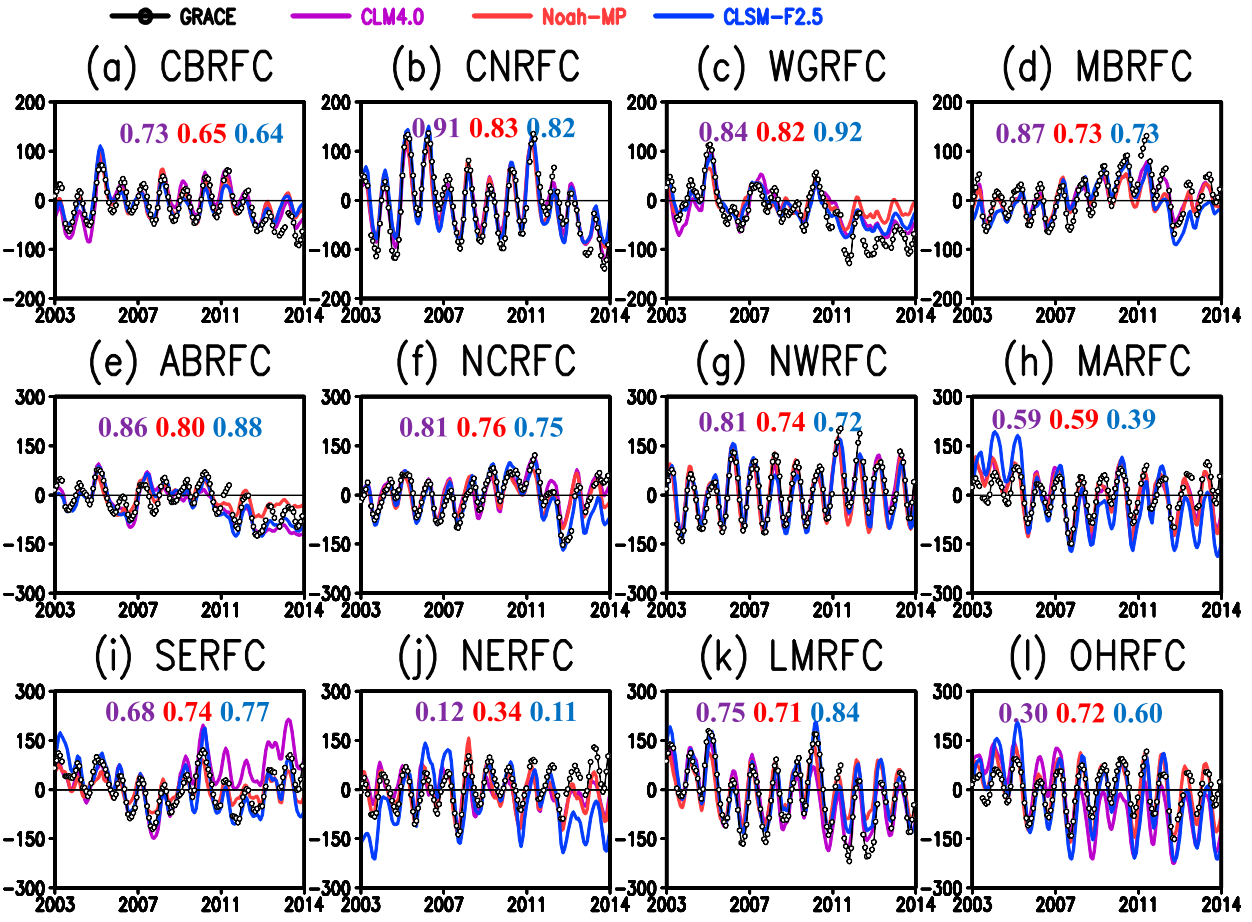


FIG. 4. Comparison of the 12-yr (2003–14) monthly time series of TWSAs (mm) from the GRACE-derived dataset (thin black line with open circles) with that simulated by the three models for 12 RFCs (purple line, CLM4.0; red line, Noah-MP; blue line, CLSM-F2.5). The TWSA climatology is calculated from January 2004 to December 2009, matching the period used by GRACE Tellus. We note that the simulated terrestrial water storage includes only total column soil moisture, SWE, and groundwater storage, thus omitting reservoir storage. The GRACE terrestrial water storage includes total column soil moisture, SWE, canopy water storage, ice, reservoir storage (e.g., rivers, lakes, and ponds), and groundwater. The basin varies from (a) the driest (CBRFC) to (l) the wettest (OHRFC) based on the aridity index represented by mean annual precipitation and potential evapotranspiration. The AC values are calculated from monthly anomaly time series with respect to the mean monthly seasonal cycle shown in Fig. 4. Values are shown for each model with the corresponding model line colors. Note that the *y*-axis ranges from −200 to +200 in (a)–(d) and from −300 to 300 in (e)–(l).

errors over wet basins than dry basins have been found in previous studies. One such study is Swenson and Lawrence (2015), which attributed the large errors as arising from large wetting events owing to the lack of a finite lower boundary in the bulk aquifer model that potentially could be fixed by removing the bulk aquifer model and adding a zero-flux boundary condition at the base of the soil column. In addition, the hydrologic parameters, including those associated with surface/subsurface runoff, soil moisture redistribution, aquifer storage and yield, and snow/snowmelt, could be tuned using data available in NLDAS to better represent the regional hydrologic regime, as demonstrated in Ren et al. (2016). The model development teams of NCAR and NASA have undertaken efforts to identify actual physical causes to reduce these large errors.

### b. *Evaluation of modeled GWSAs using USGS well data*

The first water storage component we evaluate here is groundwater storage, which is output directly from the three models for each 0.125° grid cell. For a given well, the closest grid cell is selected to compare with USGS observations at that well. For a given region, the spatial average is calculated using the same number of grid cells as USGS wells (Table 2). The modeled GWSA is calculated in the same manner as the observed anomaly from USGS wells, as described in section 3b (Li and

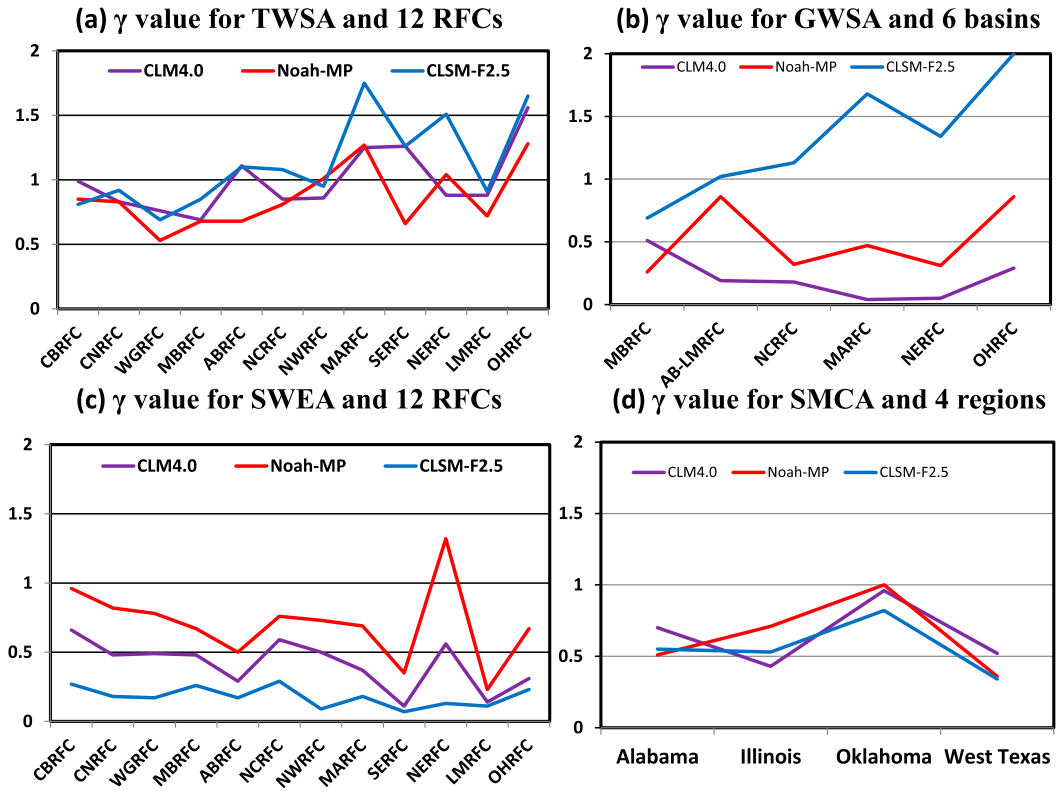


FIG. 5. The $\gamma$ value (ratio of model-simulated to GRACE-observed std dev) of the three models for (a) TWSA for 12 RFCs, (b) GWSA for 6 regions, (c) SWEA for 12 RFCs, and (d) SMCA for 4 regions (purple line, CLM4.0; red line, Noah-MP; blue line, CLSM-F2.5). The data period used here is 2003–14 for TWSA, 1980–2011 for GWSA (see Table 2), 2004–14 for SWEA, and 1980–2010 for SMCA (see Table 3).

Rodell 2015; Li et al. 2015), where the multiyear average value is subtracted from the monthly value.

Figure 7 presents the multiyear time series of monthly GWSA of the models and USGS well observations. In general, CLSM-F2.5 shows the highest GWSA anomaly amplitude (and highest anomaly bias) for all six regions, while Noah-MP shows lower anomaly amplitude and CLM4.0 shows the lowest anomaly amplitude for all six regions. The most likely reason for substantial differences among the models in GWSA magnitude is the difference in model soil depths. When precipitation variation (signal) reaches the aquifer, it is filtered by total column soil through the infiltration process. The deep (shallow) soil layer filters more (less) signal, which leads to small (large) variation magnitude in GWSA. Figure 5b shows the $\gamma$ values for GWSA. CLSM-F2.5 has the largest $\gamma$ values (higher variability than observed), while CLM4.0 has the smallest $\gamma$ values (lower variability than observed). In Fig. 6c the corresponding $E_p$ values for GWSA vary by model and basin. Overall, CLSM-F2.5 has relatively small error for all basins except for OHRFC (Fig. 6c). CLM4.0 has the deepest soil column (i.e., 3.8 m), CLSM-F2.5 has the shallowest soil column (i.e., 1 m), and Noah-MP is in between (i.e., 2 m). The recent sensitivity study of Swenson and Lawrence (2015) showed that soil layer depth has a significant impact on groundwater dynamics and storage variation. When the soil column depth in the CLM4.5 model is calibrated with GRACE TWSA (not shown), the derived optimal soil depth varies from region to region and from grid cell to grid cell. From our study we speculate that soil depth may be optimal when it is between 1 and 2 m, although further investigation is needed to examine this issue, including utilizing the newly developed global soil database (Pelletier et al. 2016),

As evident from Fig. 7, CLSM-F2.5 and Noah-MP generally yield a larger anomaly correlation for GWSA than CLM4.0, with respect to the USGS observed anomaly. All three models show very low anomaly correlations for the AB-LMRFC basin. Li and Rodell (2015) suggested that the water table in these wells may be much deeper than what these models were designed for, and the very long response time of groundwater to atmospheric forcing in drier regions like AB-LMRFC may be difficult to capture by these models. Notably,

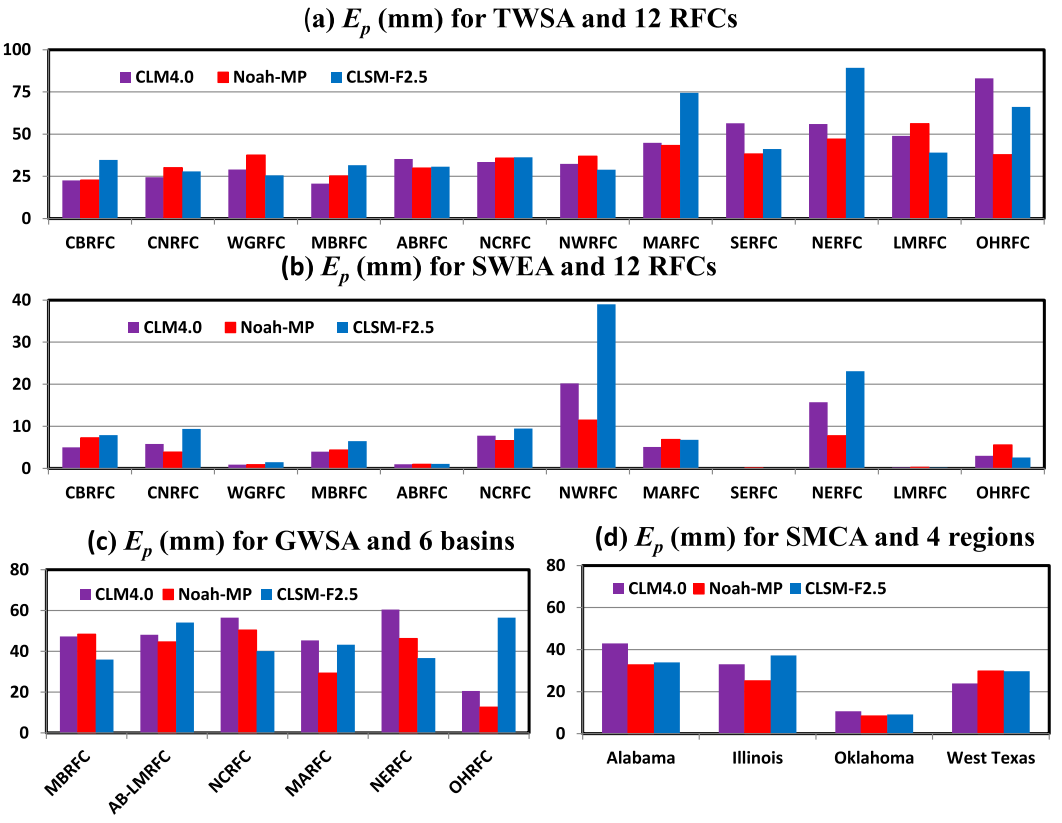


FIG. 6. The $E_p$ (mm) of the three models derived from GRACE data for (a) TWSA for 12 RFCs, (b) SWEA for 12 RFCs, (c) GWSA for 6 regions, and (d) SMCA for 4 regions (purple bar, CLM4.0; red bar, Noah-MP; blue bar, CLSM-F2.5). Note that the *y*-axis ranges from 0 to 100 in (a), from 0 to 40 in (b), and from 0 to 80 in (c),(d). The data period used here is 2003–14 for TWSA, 1980–2011 for GWSA (see Table 2), 2004–14 for SWEA, and 1980–2010 for SMCA (see Table 3).

CLM4.0 exhibits a negative anomaly correlation for NERFC, which is a problem region in which CLM4.0 has difficulty in reasonably simulating groundwater dynamics (Swenson and Lawrence 2015).

### c. *Evaluation of modeled SWEA using SNODAS*

The second water storage component we assess is SWE. In the colder regions of the United States, SWE is an important water storage term. It not only affects the estimates of terrestrial water storage but also influences total runoff/streamflow simulation and soil moisture via snowmelt. SNODAS (Clow et al. 2012) is a snow reanalysis product (see section 3c) that has been widely used to evaluate the land surface model SWE simulations and satellite retrievals (Barlage et al. 2010; Vuyovich et al. 2014). In this study, we use SNODAS as the reference analysis from which to derive model skill in terms of the AC values for the SWE simulation given in Table 4. The AC values are generally large for all three models in the four relatively cold RFCs (i.e., CNRFC, CBRFC, NWRFC, and NERFC), where snow processes including snowfall, snowmelt, and snow sublimation are substantial during the cold season, but are small in the four relatively warm RFCs (i.e., ABRFC, SERFC, LMRFC, and OHRFC), as expected. The other four RFCs have intermediate AC values. Although correlation values vary by model and by given RFC, the mean value for each model across all 12 RFCs is 0.56, 0.54, and 0.57 for CLM4.0, Noah-MP, and CLSM-F2.5, respectively, indicating a similar level of average skill across the models.

A comparison of the annual cycle of monthly mean SWE from the models and SNODAS for the 12 RFCs is shown in Fig. 8. The Noah-MP SWE simulation agrees most closely with the SNODAS over all 12 RFCs, although its amplitude is somewhat low. CLSM-F2.5 has the smallest amplitude of SWE and exhibits large underestimation compared to SNODAS (see Fig. 9). CLM4.0 also underestimates SWE relative to SNODAS, but not as much as in CLSM-F2.5. Additionally, Fig. 5c illustrates that all three models substantially underestimate ($\gamma < 1$) the temporal variability of monthly mean

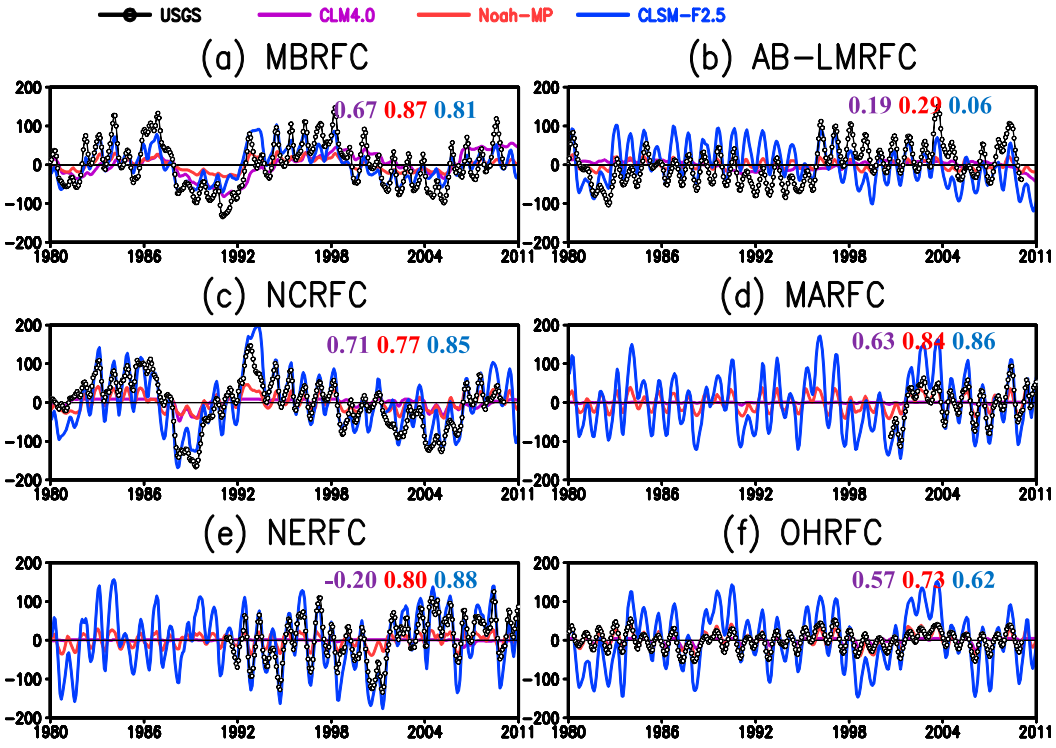


FIG. 7. Comparison of 32-yr time series of GWSA (mm) between USGS wells (thin black line with open circles) and the three models for six basins (purple line, CLM4.0; red line, Noah-MP; blue line, CLSM-F2.5). The model AC values given at the top of each frame correspond to model line color. The USGS values are unavailable before year 2002 (year 1992) for MARFC (NERFC).

SWEA compared to SNODAS in virtually all 12 RFCs, in particular CLSM-F2.5, while Noah-MP is the closest to unity ($\gamma = 1$) of the three models. Last, Fig. 6b shows that CLSM-F2.5 has the largest SWE errors (i.e., $E_p$) of the models with respect to SNODAS in the colder RFCs, especially in NWRFC and NERFC, where CLM4.0 also has large error.

We next examine an example of midwinter spatial distribution of SWE across the continental United States for SNODAS and the three models in Fig. 9, which depicts the CONUS distribution of 11-yr-average SWE for February. Noah-MP exhibits results comparable to SNODAS SWE in the western, north-central, and northeastern regions. All three models exhibit a spatial pattern of SWE coverage comparable to that of SNODAS, except Noah-MP and CLM4.0 somewhat overestimate the coverage in the southeastern and south-central United States. More significantly, all three models significantly underestimate the SWE magnitude compared to SNODAS in the higher elevations of the western United States, though Noah-MP shows less underestimation than CLM4.0 and CLSM-F2.5 (SNODAS may slightly overestimate monthly SWE observations, in particular at low elevations; Clow et al. 2012). CLSM-F2.5 in particular substantially underestimates the SWE magnitudes of SNODAS over the western highest elevations.

To investigate why CLSM-F2.5 generates notably low SWE, we compare the snowfall, snowmelt, and snow sublimation output of the CLSM-F2.5 and Noah-MP. First, we use 35 years (1980–2014) of model output to calculate the average annual cycle of each model to ensure a reliable climatology for each model. The snow models used in Noah-MP and CLSM-F2.5, albeit different models, both incorporate extensive snowpack physics, including evaporation, sublimation, condensation, radiation, precipitation as rain or snowfall, mechanical

TABLE 4. Coefficients of monthly AC between basinwide averaged SWEA derived from SNODAS and three models for 12 RFCs.

| RFC name | CBRFC | CNRFC | WGRFC | MBRFC | ABRFC | NCRFC | NWRFC | MARFC | SERFC | NERFC | LMRFC | OHRFC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CLM4.0 | 0.70 | 0.79 | 0.60 | 0.61 | 0.20 | 0.80 | 0.85 | 0.66 | 0.17 | 0.73 | 0.20 | 0.45 |
| Noah-MP | 0.67 | 0.79 | 0.62 | 0.58 | 0.25 | 0.80 | 0.75 | 0.55 | 0.13 | 0.70 | 0.13 | 0.47 |
| CLSM-F2.5 | 0.68 | 0.79 | 0.54 | 0.62 | 0.36 | 0.78 | 0.61 | 0.81 | 0.22 | 0.63 | 0.23 | 0.62 |

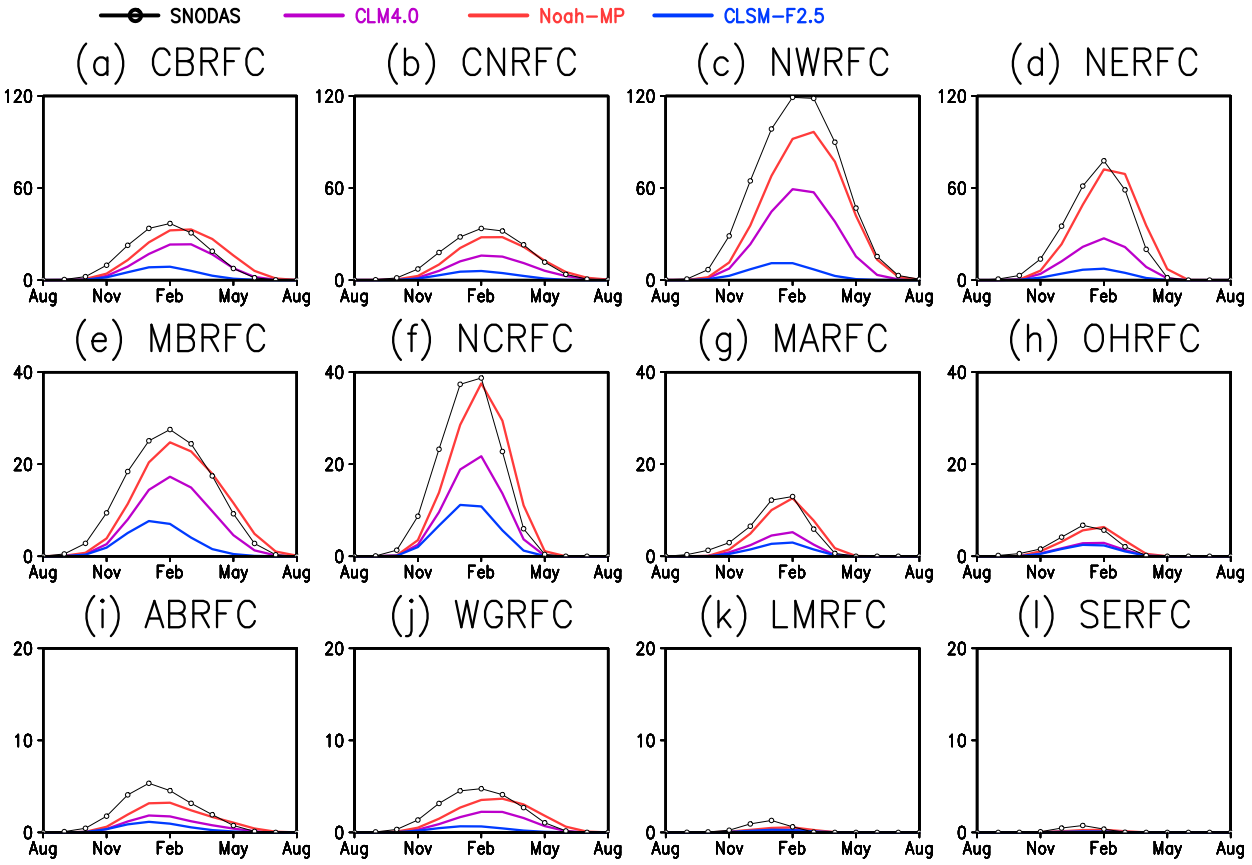


FIG. 8. Mean annual cycle of monthly mean SWE (mm) for each of 12 RFCs for the three models and SNODAS reanalysis calculated for the 11-yr period of 2004–14 (thin black line with open circles, SNODAS; purple line, CLM4.0; red line, Noah-MP; blue line, CLSM-F2.5). Note that the *y*-axis ranges from 0 to 120 in (a)–(d), from 0 to 40 in (e)–(h), and from 0 to 20 in (i)–(l).

compression, overflow, underflow, snowmelt, etc. (Niu et al. 2011; Lynch-Stieglitz 1994; Stieglitz et al. 2001). Such complicated snowpack models make it difficult to diagnose/identify SWE bias problems. In the case of Noah-MP, Niu et al. (2011) compared the Noah-MP to the earlier Noah, version 3 (with its single bulk-layer snowpack model), to elucidate Noah-MP's advantage in simulating SWE. Major findings with Noah-MP were that its multilayer snowpack structure with the liquid water retention, optimal surface turbulent exchange coefficient, and snow surface roughness length (i.e., 0.002 m) play an important role in simulating SWE.

Here we introduce a bulk snowpack model concept to guide a diagnosis of why CLSM-F2.5 produces notably small SWE. In general terms, the temporal change of SWE is the difference of the input snowfall (source) minus both the snow sublimation/evaporation and the snowmelt (sinks), which can be expressed as

$$\frac{dW_s}{dt} = P_s - M_s - E_s, \qquad (2)$$

where $W_s$ denotes SWE, $t$ is time, $E_s$ is the rate of sublimation/evaporation, $P_s$ is snowfall rate, and $M_s$ is snowmelt rate, which is controlled by the net energy input to snowpack, according to

$$M_s = \frac{1}{L}(Q_{\mathrm{sw}} + Q_{\mathrm{lw}} - Q_{\mathrm{sh}} - Q_{\mathrm{le}} - Q_g), \qquad (3)$$

where $Q_{\mathrm{sw}}$ is net solar (shortwave) radiation, $Q_{\mathrm{lw}}$ is net longwave radiation, $Q_{\mathrm{le}}$ is the latent heat flux, $Q_{\mathrm{sh}}$ is the sensible heat flux, $Q_g$ is ground heat flux, and $L$ is the latent heat of fusion.

We selected four RFCs where snowpack is important in winter to compare in Fig. 10 the annual cycle of the monthly values of the three terms on the right side of Eq. (2) for both CLSM-F2.5 and Noah-MP. Compared to Noah-MP, Fig. 10 shows that in CLSM-F2.5 the sublimation sink is clearly larger, while its snowmelt sink is simultaneously lower by roughly the same magnitude (except in the MBRFC. Note in Fig. 10 that the *y*-axis range in Figs. 10i–l is one-fourth the range of Figs. 10a–h). These similar magnitudes but opposite sign of the

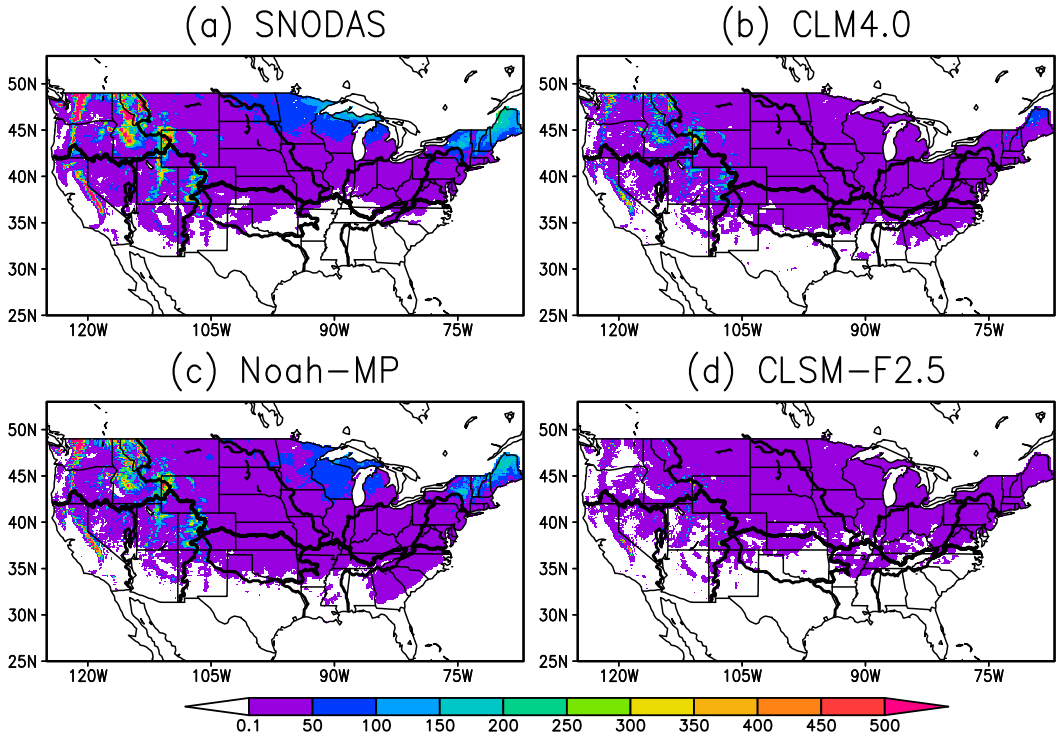


FIG. 9. Spatial distribution of 11-yr (2004–14) averaged February SWE (mm) derived from (a) SNODAS, (b) CLM4.0, (c) Noah-MP, and (d) CLSM-F2.5.

sublimation and snowmelt differences bring our attention to snowfall as the likely major reason for the large differences in SWE between CLSM-F2.5 and Noah-MP. There is substantially lower snowfall in CLSM-F2.5 than in Noah-MP (Figs. 10a–d) in three of the four RFCs, as well as slightly lower snowfall in the fourth RFC, despite the fact that both models receive the same hourly forcing of total precipitation. Thus, the two models partition total precipitation differently into rainfall and snowfall, which we examine further next.

Noah-MP uses Jordan's scheme (Jordan 1991) to fractionally divide total precipitation into snowfall and rainfall, while CLSM-F2.5 uses a simple 0.0°C threshold of air temperature to specify either all rainfall or all snowfall. Of the input total precipitation, the fraction of frozen precipitation $F_{ice}$ in Jordan's scheme can vary within the range of 0–1 as determined by

$$F_{ice} = \begin{cases} 0 & T_{air} > 2.5^{\circ}\text{C} \\ 0.6 & 2.0^{\circ} < T_{air} \le 2.5^{\circ}\text{C} \\ 1 - [54.62 - 0.2(T_{air} + 273.15)] & 0.0^{\circ} < T_{air} \le 2.0^{\circ}\text{C} \\ 1.0 & T_{air} \le 0.0^{\circ}\text{C} \end{cases}, \quad (4)$$

In CLSM-F2.5, the fraction of frozen precipitation is specified in a binary fashion as

$$F_{ice} = \begin{cases} 0 & T_{air} > 0.0^{\circ}\text{C} \\ 1 & T_{air} \le 0.0^{\circ}\text{C} \end{cases}, \quad (5)$$

where $T_{air}$ is the 2-m air temperature. Thus, in CLSM-F2.5, all of the precipitation falls as rain unless the 2-m air temperature is 0.0°C or colder, whereas Noah-MP applies Jordan's scheme that utilizes a warmer 2.5°C air temperature criterion for the onset of at least some frozen precipitation (which includes snow, snow mixed with frozen rain, and frozen rain), leading to more solid precipitation in the Noah-MP.

We now further consider the large difference in sublimation between CLSM-F2.5 and Noah-MP in all four RFCs in Figs. 10i–l. As a point of reference, when the modeled snowmelt (Figs. 10e–h) is scaled with each model's own mean annual snowfall (not shown), their fractions (snowmelt/snowfall) are similar, suggesting

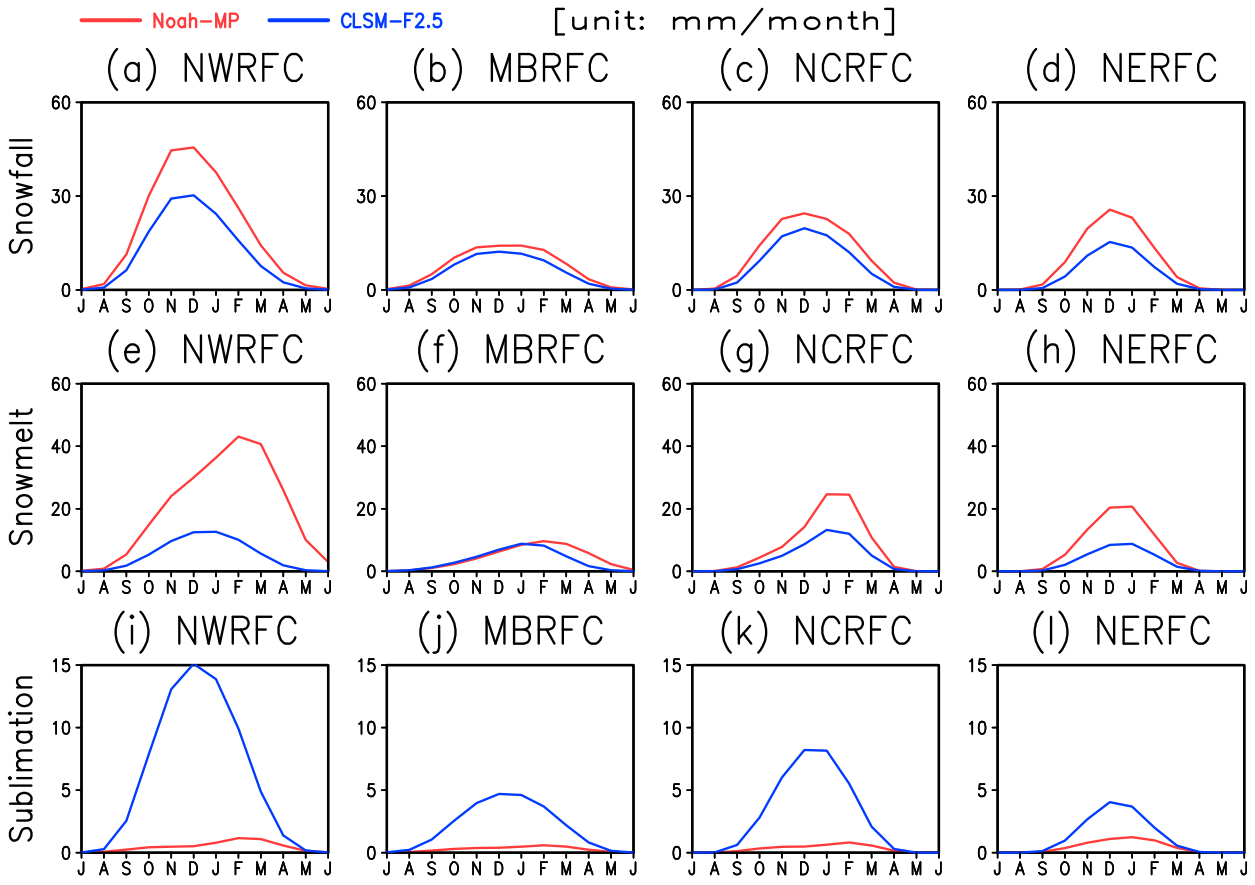


FIG. 10. Mean annual cycle of monthly mean snowfall, snowmelt, and sublimation (mm) for four selected RFCs for CLSM-F2.5 (blue line) and Noah-MP (red line) calculated from 35-yr (1980–2014) monthly model output. Note that the *y*-axis is from 0 to 60 mm in (a)–(h) and from 0 to 15 mm in (i)–(l).

that the snowmelt differences between the two models are of less importance than the large sublimation differences in explaining the large differences in SWE. The sublimation/evaporation over a snow surface can be simply represented by the following bulk transfer formula:

$$E_s = F_{sc} C_h U (q_s^* - q_0), \tag{6}$$

where $F_{sc}$ is the snow cover fraction associated with SWE and snow density, $U$ is mean wind speed at 10-m height, $C_h$ is the surface exchange coefficient, $q_s^*$ is the surface saturated specific humidity, and $q_0$ is the atmospheric specific humidity at 2-m height. The surface exchange coefficients are calculated as functions of snow surface roughness and an atmospheric stability parameter, which is the Monin–Obukhov length in Noah-MP and the bulk Richardson number in CLSM-F2.5. Given that the input $U$ and $q_0$ are the same in both models, and given that the snow cover fraction will be close to 1 in both models when SWE is nontrivial, the major difference in sublimation between CLSM-F2.5 and Noah-MP must arise from their differences in $C_h$ and $q_s^*$, which we leave for a future study.

### d. *Evaluation of modeled SMCA using NASMD*

Last, the third water storage component we evaluate is soil moisture. Like groundwater storage observations, soil moisture measurements at the deeper depths are difficult to obtain. Therefore, evaluating modeled soil moisture is a very challenging task. Although the NASMD project compiled more than 1800 stations from 30 soil moisture measurement networks across the United States, there are few stations that measure soil moisture below the 2-m soil depth, as most stations measure soil moisture near the land surface (Quiring et al. 2016). Furthermore, many stations located in mountainous/cold regions (e.g., SNOTEL network) have over 70% missing records and/or have erroneous records for most of the cold season (September–May), when frozen soil may lead to instrument malfunctions (Xia et al. 2015c). We use the three criteria to select soil

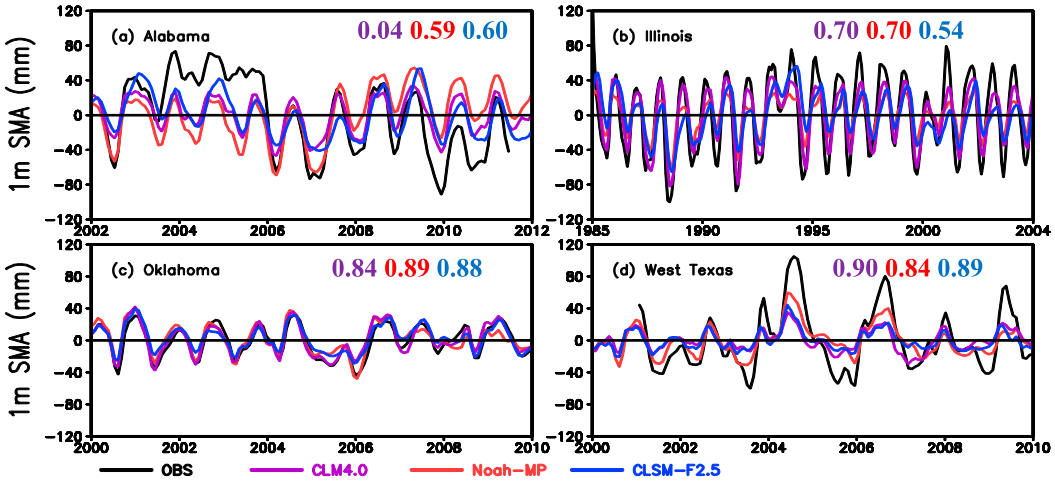


FIG. 11. Comparison of multiyear (from 11 to 20 years) time series of top 1-m monthly mean SMCA (mm) spatially averaged over the NASMD sites for each of four regions from the NASMD observations (thin black line) and the three models, along with the model AC values (purple line, CLM4.0; red line, Noah-MP; blue line, CLSM-F2.5). The number of NASMD sites averaged in each region is listed in Table 3.

moisture measurement stations (Xia et al. 2015a,c) as follows. First, there are at least 10 years of observations; second, the observations span at least a complete year; and third, the depth to the soil measurement is at least 1 m or deeper. Based on these criteria, we selected the four regions of Alabama, Illinois, Oklahoma, and western Texas (Table 3).

As all three models and the four regions mentioned above contain the top 1-m soil moisture, we used the 195 sites given in Table 3 and plotted in Fig. 2 to evaluate the modeled soil moisture. To minimize the effect of scale incompatibility problems on our soil moisture evaluation (Entin et al. 2000; Robock et al. 2003; Fan et al. 2006; Xia et al. 2014), we used spatial averages for both modeled and observed soil moisture. Spatially averaged soil moisture anomalies from NASMD and the three models are shown in Fig. 11 for each of the four regions mentioned above. Overall, the models reasonably capture monthly variability in the observed soil moisture anomaly in the four regions, in particular Oklahoma and Illinois, while the amplitude is smaller than the observations in Alabama and western Texas, where the models tend to underestimate the larger amplitude anomalies of the several wettest and driest periods. The anomaly correlation coefficients (ACCs) are relatively high for all models and regions (>0.66) except for CLM4.0 in Alabama, which shows a very low anomaly correlation. For the larger regions such as the four quadrants of the United States, Cai et al. (2014b) used 121 stations from the Soil and Climate Analysis Network (SCAN) to evaluate the top 1-m soil moisture simulated from CLM4.0 and Noah-MP. The results in Fig. 11 are in good agreement with their study. Additionally, in our study here, all three models underestimate SMCA temporal variability (Fig. 5d) and exhibit similar moderate errors (Fig. 6d) in all four regions.

We recognize that this evaluation is limited in the case of CLM4.0 and Noah-MP, which have total soil column depths of 3.8 and 2.0 m, respectively. However, only one of the four regions (Illinois) has a meaningful number of soil moisture observations below 1 m. For the sake of brevity, we leave evaluation of deeper soil moisture for a future study, for which more widespread observations of deeper soil moisture may become available.

## 5. Analysis of relative contribution of different water storage components

Although terrestrial water storage anomalies from the three models are relatively similar and comparable to GRACE TWSA for the 12 RFCs (see section 4a), their individual components are quite different. Comparison and evaluation of groundwater storage and SWE have shown large intermodel differences. Although the top 1-m soil moisture does not show large differences between models, the model SMCAs in the total soil column have large differences because of different soil depths; as expected, CLM4.0 has the highest amplitude, followed by Noah-MP and CLSM-F2.5 (Fig. 6d). An analysis for mean monthly variability and interannual variability can reveal more about the relative contribution of each component to the terrestrial water storage by each model. Thus, we calculate the mean annual cycle and annual mean anomalies for the 35-yr (1980–2014) monthly anomalies at each RFC. The relative contribution from

the mean annual cycle $RC_{monthly}$ and interannual variability $RC_{interann}$ is calculated as

$$RC_{monthly} = \frac{\sigma_i}{\sigma_T} \quad \text{and} \tag{7a}$$

$$RC_{interann} = \frac{\delta_i}{\delta_T}, \tag{7b}$$

where $\sigma_i$ is the standard deviation of the monthly anomaly of the $i$th water storage component and $\sigma_T$ is the standard deviation of the monthly anomaly of the terrestrial water storage, both calculated with respect to the 35-yr mean seasonal cycle. We also calculate the standard deviation for the annual anomaly of the $i$th water component $\delta_i$ and the standard deviation of the annual anomaly of the terrestrial water storage $\delta_T$, both calculated with respect to their 35-yr (1980–2014) average value.

The relative contributions of the three storage components (SMCA, SWEA, and GWSA) are normalized to sum to one and plotted on a ternary diagram for the mean monthly variability (Figs. 12a–c) and interannual variability (Figs. 12d–f). The diagram has three apexes representing the three components, and the distance from each data point to each apex reflects the proportion of that component within the TWSA. For the mean monthly variability (Figs. 12a–c), the major contributor to TWSA is SMCA for CLM4.0 and Noah-MP and GWSA for CLSM-F2.5. When CLM4.0 is compared with Noah-MP, SMCA contributes somewhat more and GWSA contributes somewhat less, as its deeper soil layers yield larger SMCA amplitude and weaker GWSA amplitude. The small SWEA contribution to TWSA apparent in CLSM-F2.5 is due to its low SWE simulations, as expected.

The relative contribution of GWSA to TWSA increases for all three models when interannual variability is considered (Figs. 12d–f), particularly for CLM4.0. As groundwater storage has long memory and large interannual variability, its contribution becomes larger when compared to its contribution to the monthly variability. This is consistent with the earlier study of Güntner et al. (2007). In CLSM-F2.5, the contributions to interannual variability from SWEA and SMCA become much smaller compared to their contributions to monthly variability, due in part to the shallower soil (1 m) in CLSM-F2.5.

Studies using in situ observations in Illinois and Oklahoma (Swenson et al. 2006, 2008) showed that SMCA and GWSA have a similar magnitude and amplitude. The difference between our result and that analysis is caused by model structure (e.g., soil layer depth) and model simulation errors (e.g., SWE and groundwater storage), which are usually closely related. These issues highlight the need to improve model physics and to establish networks of observations for all water storage components (e.g., soil moisture, wells, and SWE) to benefit model development.

## 6. Discussion

Overall results show that our model skills in simulating groundwater are very limited in most regions within the United States without more work on 1) a better understanding of the interplay between model structure, model parameters, and model setup; 2) a better understanding of groundwater dynamics such as the impact of soil depth and aquifer depth and lateral connectivity between the grid cells; and 3) quantifying the effects of soil water, surface runoff, and baseflow recharge on groundwater. The intermodel differences of groundwater storage are quite large, in particular the different anomaly amplitudes generated by the three models.

Persistence in TWSA can be quantified by the autocorrelation of the ACC. Figure 13 shows the ACC values observed from GRACE and simulated from the three models at four RFCs as a function of the starting month and target (lag) month. For the four RFCs of CBRFC, MBRFC, SERFC, and OHRFC, the first two are located in drier regions and the second two are located in wetter regions based on the aridity index (ratio of mean annual precipitation to potential evapotranspiration) as used in Sankarasubramanian and Vogel (2002). Overall, the model results agree reasonably well with observations in their cyclic phases. For the observations, a significant lag correlation at the 95% confidence level persists 1–6 months depending on different RFCs and months [Fig. 13 (top)]. Generally, CLM4.0 has the highest persistence when compared with the other two models at all four RFCs. In the relatively dry CBRFC and MBRFC, CLM4.0 is rather close to the observations, while Noah-MP and CLSM-F2.5 both underestimate the persistence in CBRFC. In the wetter SERFC and OHRFC, CLSM-F2.5 is relatively close to the observed, while Noah-MP still exhibits underestimation. A stronger persistence suggests that a dry or wet event can take a longer time to recover to its normal condition. This means that a drought event represented by TWSA in CLM4.0 (Noah-MP) may persist longer (shorter) than expected from the observation (GRACE), indicating a longer (shorter) duration and later (earlier) termination of a given drought event. Given these differences, when TWSA is used to monitor drought events, this would have important implications for interpretation and utilization of the model simulations. A companion analysis of the persistence of the simulated SMCA and GWSA storage components in Fig. 14 shows that CLM4.0 has the largest

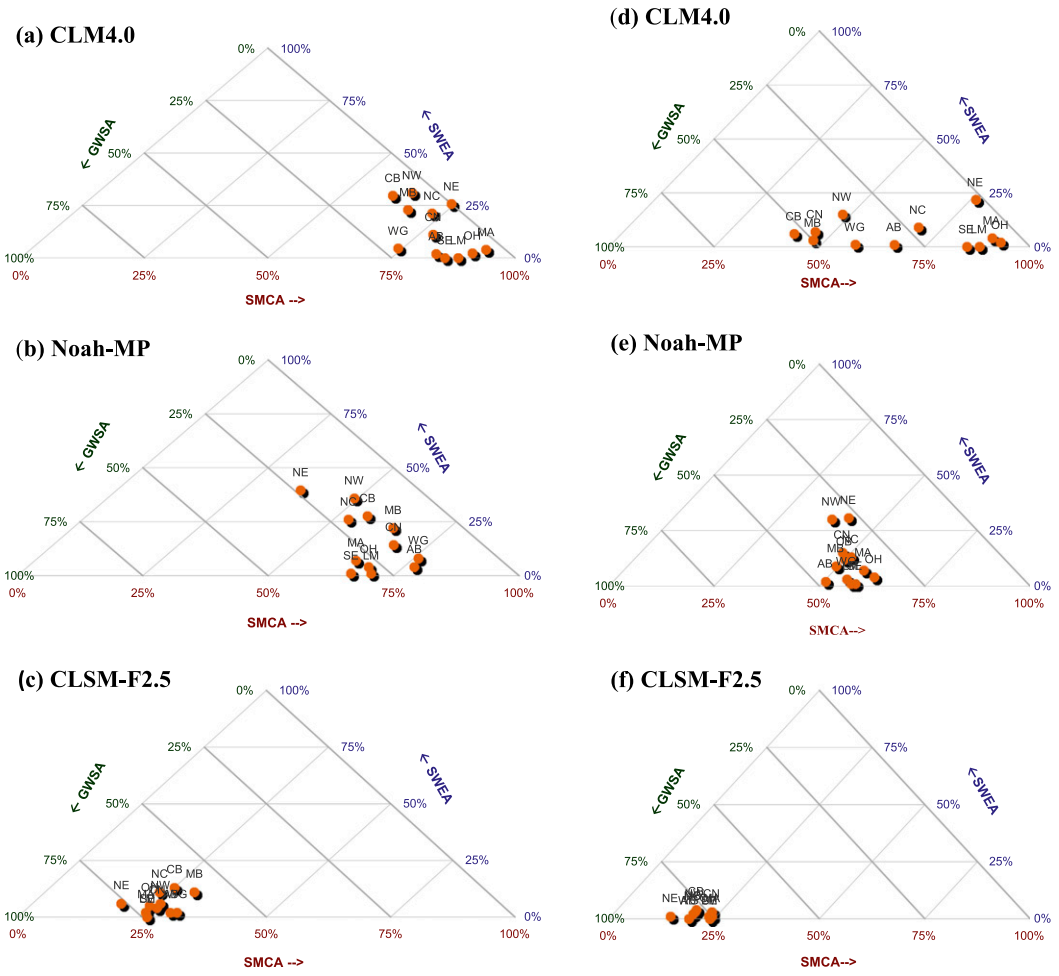


FIG. 12. Ternary plot showing the relative contributions of the three water storage terms (SMCA, GWSA, and SWEA) calculated from 35-yr (1980–2014) model output for the 12 RFCs within the CONUS. The storage terms for each basin were normalized to the sum of one. (a)–(c) Mean monthly variability and (d)–(f) interannual variability are shown.

ACC values for both SMCA and GWSA when compared with the other two models. For Noah-MP and CLSM-F2.5, the ACC values are comparable for SMCA while the CLSM-F2.5 has the smaller ACC values for GWSA.

We point out here that besides the assessment of TWSA, SWEA, SMCA, and GWSA in this present paper, we have completed important companion evaluations of monthly streamflow and evapotranspiration against observations for two of these three models (Cai et al. 2014b; Xia et al. 2016). The observed streamflow from 981 USGS small- to medium-sized basins ($<10\,000\,km^2$) were used to validate CLM4.0, Noah-MP, Noah, and VIC. The latter two models are being used in the current operational NLDAS. The results showed that Noah-MP and VIC have larger AC values and smaller biases when compared to CLM4.0 and the older Noah model (Cai et al. 2014b). However, overall results were comparable for the four models. CLSM-F2.5 and the Mosaic model (which is used in the current operational NLDAS) were also evaluated against monthly gridded USGS total runoff and gridded FLUXNET evapotranspiration generated from Jung et al. (2009) at 12 RFCs (Xia et al. 2016). The results showed that CLSM-F2.5 has much smaller (larger) total runoff (evapotranspiration) when compared to the observations and the Mosaic model. This problem is being addressed in the NLDAS Science Test bed. This ongoing research work will help our NLDAS team to decide if the Mosaic model will be replaced by CLSM-F2.5 in the next-generation operational NLDAS.

## 7. Conclusions

This paper assessed the total (terrestrial) water storage components of three advanced LSMs (CLM4.0,

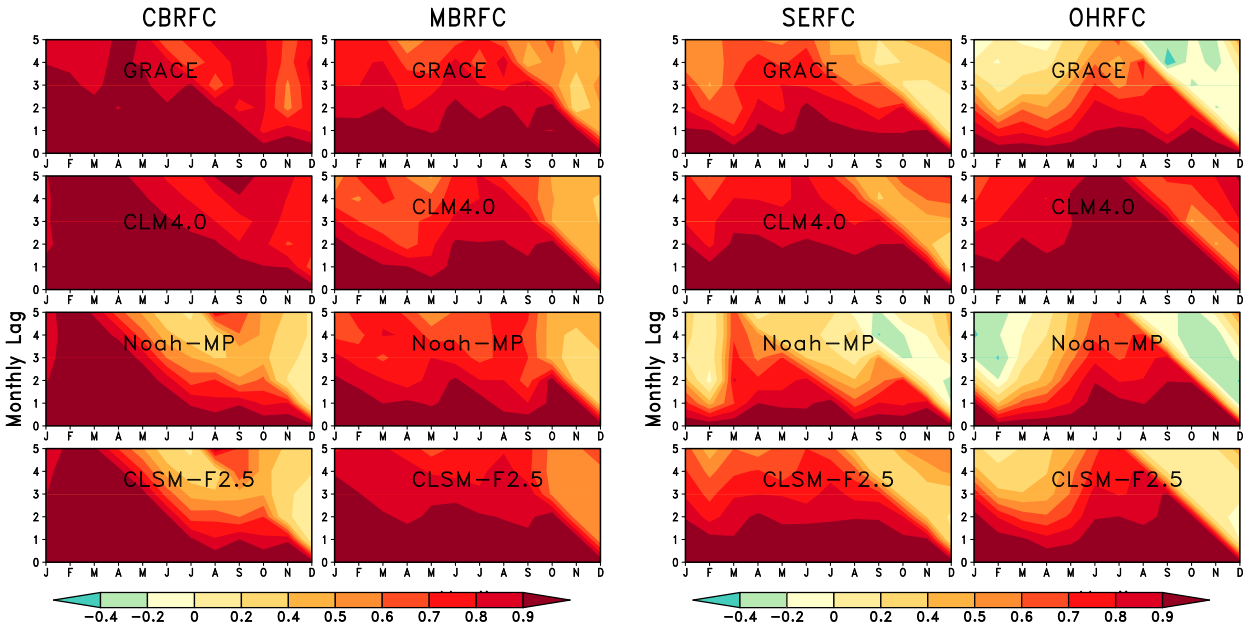


FIG. 13. Seasonal dependence of observed and simulated TWSA persistence (month to month autocorrelation) for 2003–14 over the four RFCs (CBRFC, MBRFC, SERFC, and OHRFC). Initial month is along the *x* axis, and the lag to the target month is along the *y* axis. Observations from GRACE and three models (CLM4.0, Noah-MP, and CLSM-F2.5) are shown from top to bottom, and the results from the four RFCS are shown from left to right.

Noah-MP, and CLSM-F2.5). The validation of the monthly TWSA against GRACE-based observations for 12 RFCs showed that all three LSMs capture the broad features of satellite-observed TWSA variations, such as dry and wet events, although the LSMs' simulation skill varies from model to model and basin to basin. The modeled mean annual cycle of TWSA agrees rather well with GRACE observations for almost all RFCs.

The comparison of modeled groundwater storage with USGS well observations shows that overall simulation skill (anomaly correlation) is consistent with previous studies (Li and Rodell 2015; Cai et al. 2014a), although CLM4.0 fails to capture the GWSA variation for the NERFC. All models have a low anomaly correlation in the combined AB-LMRFC basin when compared to the other basins, likely due to the deeper in situ wells and drier climate, which makes it harder for the models to capture groundwater response time. Other factors such as irrigation/pumping (Leng et al. 2014) that are not considered by all three models may also play a role but cannot be verified at this time.

However, in contrast to temporal variability of monthly GWSA, the magnitudes of monthly GWSA have large intermodel differences in all six basins evaluated. In general, CLM4.0 and Noah-MP have lower amplitudes of monthly GWSA and CLSM-F2.5 has higher amplitudes when compared to the USGS observations. The most likely reason is that different model structures (soil layer depth and unconfined aquifer depth) lead to these differences. For instance, CLM4.0 has deeper soil layer that may lead to higher soil moisture variability than in the other models such as CLSM-F2.5.

An evaluation of modeled SWE against SNODAS products shows that Noah-MP is the closest to SNODAS, followed by CLM4.0 and CLSM-F2.5. However, CLSM-F2.5 largely underestimates the reference SWE when compared to SNODAS. A major reason is the large sublimation rates of CLSM-F2.5, similar to those found in the earlier Noah LSM versions preceding the Noah-MP (Slater et al. 2007; Livneh et al. 2010). In addition, the colder air temperature threshold used in CLSM-F2.5 to partition the input of total precipitation into snowfall and rainfall is a major factor in the low bias of SWE in CLSM-F2.5.

The soil moisture evaluation in this study is somewhat narrow in its scope. Specifically, although the NASMD collected over 1800 stations across the United States, only 195 stations are used here owing to three criteria we applied for choosing validating soil moisture observations. The resulting validation of the LSMs against the observations of the top 1-m soil moisture shows that all models can reasonably capture soil moisture anomalies such as dry and wet events with a relatively high skill. There are large differences among the three LSMs in the amplitude of the soil moisture anomalies in Alabama and western Texas, while all three LSMs exhibit better

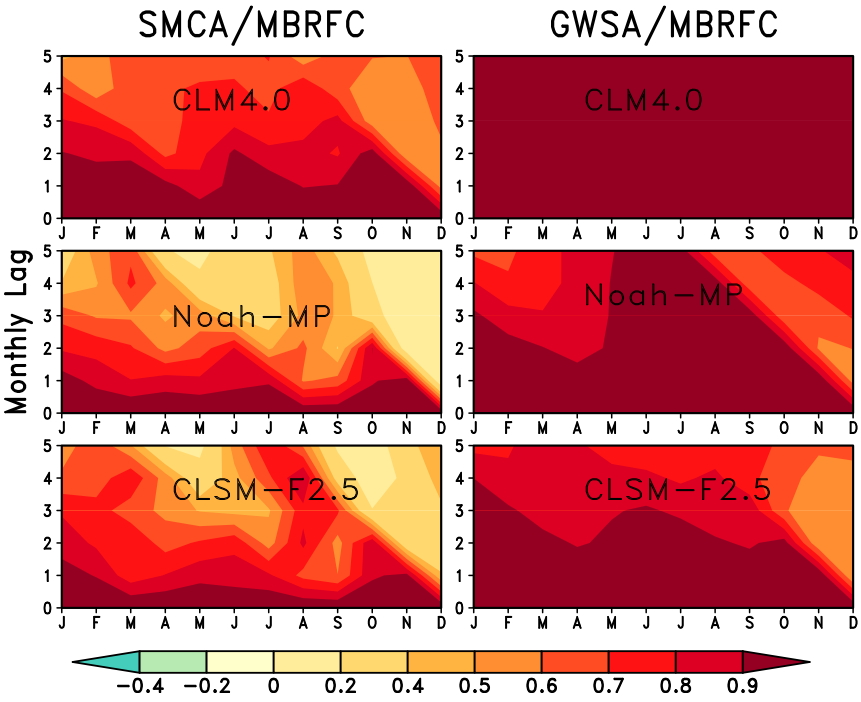


FIG. 14. Seasonal dependence of simulated (left) SMCA and (right) GWSA persistence (month to month autocorrelation) for 2003–14 at MBRFC for three models (CLM4.0, Noah-MP, and CLSM-F2.5). Initial month is along the *x* axis, and the lag to the target month is along the *y* axis. Observations from GRACE and three models (CLM4.0, Noah-MP, and CLSM-F2.5) are shown from top to bottom.

performance in Oklahoma and Illinois. However, because of a lack of deeper soil moisture measurements, a comprehensive evaluation is still lacking and further validation is needed when such measurements become available.

We presented an analysis of the relative contribution of individual water storage components anomaly to TWSA. Over the 12 RCFs over the CONUS, soil moisture is a major contributor to TWSA for CLM4.0 and Noah-MP, while groundwater dominates TWSA in CLSM-F2.5. We emphasize that such an analysis is preliminary because of quite different model structures (e.g., different soil layer numbers and depth and aquifer depth), model deficiencies and errors (e.g., low SWE simulation in CLSM-F2.5), and individual water component simulation errors. However, such large intermodel differences suggest that further investigation is needed.

Based on this study, we recommend future work in the following areas: 1) scientific assessment of model designs such as soil layer thicknesses and water-table depth, interaction between groundwater and surface water, and groundwater response time to precipitation; 2) improved model parameters and calibration techniques that are suitable for large-scale modeling; 3) addition of other realistic physical processes such as irrigation/pumping and surface water storage (ponds, lakes, rivers, and wetlands); and 4) additional networks of in situ and satellite-based observations to facilitate model improvements.

As mentioned in our introduction, there has been quite a bit of work on the topic of land surface models versus GRACE comparisons. Some of this work may reach slightly different conclusions than those presented here. For example, van Dijk et al. (2014) indicated that land surface models can reasonably capture interannual trends in terrestrial water storage. For the most part, this conclusion is consistent with ours. The three models used in this study capture TWSA reasonably well when compared with GRACE TWSA. However, for each individual component, different models show different performances. In particular for GWSA, the simulation skills are very limited for all three models based on our preliminary evaluation. As more well data become available regionally and globally, a further validation for these three models as well as the other models with a groundwater module can reach more robust conclusions in the future.

*Acknowledgments.* Y.X. at NCEP/Environmental Modeling Center and D.M. at NASA GSFC Hydrological Sciences Laboratory were partially sponsored by NOAA Climate Program Office (CPO) Modeling,

Analysis, Predictions, and Projections (MAPP) program (Climate Program Office Award Number GC14-194A). M.H. is supported by the Subsurface Biogeochemical Research (SBR) program through the Pacific Northwest National Laboratory (PNNL) SBR Scientific Focus Area (SFA). PNNL is operated by Battelle Memorial Institute for the U.S. Department of Energy under contract DE-AC05-76RLO1830. B.L. at the University of Maryland, College Park, and M.R. at NASA GFSC Hydrological Sciences Laboratory were sponsored by NASA's Terrestrial Hydrology Program. We thank Mary Hart and Jeff McQueen at the Environmental Modeling Center (EMC) and three anonymous reviewers whose comments greatly improved readability and quality of our manuscript.

## REFERENCES

Alexeev, V. A., D. J. Nicolsky, V. E. Romanovsky, and D. M. Lawrence, 2007: An evaluation of deep soil configurations in the CLM3 for improved representation of permafrost. *Geophys. Res. Lett.*, **34**, L09502, doi:10.1029/2007GL029536.

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, doi:10.5194/hess-17-5127-2013.

Azar, A. E., H. Ghedira, P. Romanov, S. Mahani, M. Tedesco, and R. Khanbilvardi, 2008: Application of satellite microwave images in estimating snow water equivalent. *J. Amer. Water Resour. Assoc.*, **44**, 1347–1362, doi:10.1111/j.1752-1688.2008.00227.x.

Ball, J. T., I. E. Woodrow, and J. A. Berry, 1987: A model predicting stomatal conductance and its contribution to the control of photosynthesis under different environmental conditions. *Progress in Photosynthesis Research*, Vol. 4, J. Biggins, Ed., Springer, 221–224, doi:10.1007/978-94-017-0519-6_48.

Barlage, M., and Coauthors, 2010: Noah land surface model modifications to improve snowpack prediction in the Colorado Rocky Mountains. *J. Geophys. Res.*, **115**, D22101, doi:10.1029/2009JD013470.

——, M. Tewari, F. Chen, C. Migue-Macho, Z.-L. Yang, and G.-Y. Niu, 2015: The effect of groundwater interaction in North American regional climate simulations with WRF/Noah-MP. *Climatic Change*, **129**, 485–498, doi:10.1007/s10584-014-1308-8.

Bell, J. E., and Coauthors, 2013: U.S. Climate Reference Network soil moisture and temperature observations. *J. Hydrometeor.*, **14**, 977–988, doi:10.1175/JHM-D-12-0146.1.

Berg, A. A., J. S. Famiglietti, J. P. Walker, and P. R. Houser, 2003: Impact of bias correction to reanalysis products on simulations of North American soil moisture and hydrological fluxes. *J. Geophys. Res.*, **108**, 4490, doi:10.1029/2002JD003334.

Bingham, R. J., and C. W. Hughes, 2006: Observing seasonal bottom pressure variability in the North Pacific with GRACE. *Geophys. Res. Lett.*, **33**, L08607, doi:10.1029/2005GL025489.

Bonan, G. B., 1996: A land surface model (LSM version 1.0) for ecological, hydrological, and atmospheric studies: Technical description and user's guide. NCAR Tech. Note NCAR/TN-417+STR, 150 pp., doi:10.5065/D6DF6P5X.

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

Cai, X. T., Z.-L. Yang, C. H. David, G.-Y. Niu, and M. Rodell, 2014a: Hydrological evaluation of the Noah-MP land surface model for the Mississippi River basin. *J. Geophys. Res. Atmos.*, **119**, 23–38, doi:10.1002/2013JD020792.

——, ——, Y. Xia, M. Huang, H. Wei, R. Leung, and M. B. Ek, 2014b: 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, doi:10.1002/2014JD022113.

Carroll, T., D. Cline, C. Olheiser, A. Rost, A. Nilsson, G. Fall, C. Bovitz, and L. Li, 2006: NOAA's national snow analyses. *Proc. Western Snow Conf.*, Vol. 74, Las Cruces, NM, Hydrological Services America, 31–43. [Available online at www.westernsnowconference.org/sites/westernsnowconference.org/PDFs/2006Carroll.pdf.]

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, doi:10.1002/hyp.9385.

Cosby, B. J., G. M. Hornberger, R. B. Clapp, and T. R. Ginn, 1984: A statistical exploration of the relationships of soil moisture characteristics to the physical properties of soils. *Water Resour. Res.*, **20**, 682–690, doi:10.1029/WR020i006p00682.

Cosgrove, B. A., and Coauthors, 2003a: Real-time and retrospective forcing in the North American Land Data Assimilation System (NLDAS) project. *J. Geophys. Res.*, **108**, 8842, doi:10.1029/2002JD003118.

——, and Coauthors, 2003b: Land surface model spin-up behavior in the North American Land Data Assimilation System (NLDAS). *J. Geophys. Res.*, **108**, 8845, doi:10.1029/2002JD003316.

Csiszar, I., and G. Gutman, 1999: Mapping global land surface albedo from NOAA AVHRR. *J. Geophys. Res.*, **104**, 6215–6228, doi:10.1029/1998JD200090.

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, doi:10.1175/1520-0450(1994)033<0140:ASTMFM>2.0.CO;2.

De Lannoy, G. J. M., and R. H. Reichle, 2016: Global assimilation of multiangle and multipolarization SMOS brightness temperature observations into the GEOS-5 catchment land surface model for soil moisture estimation. *J. Hydrometeor.*, **17**, 669–691, doi:10.1175/JHM-D-15-0037.1.

——, ——, and V. R. N. Pauwels, 2013: Global calibration of the GEOS-5 L-band microwave radiative transfer model over land using SMOS observations. *J. Hydrometeor.*, **14**, 765–785, doi:10.1175/JHM-D-12-092.1.

Dickinson, R. E., M. Shaikh, R. Bryant, and L. Graumlich, 1998: Interactive canopies for a climate model. *J. Climate*, **11**, 2823–2836, doi:10.1175/1520-0442(1998)011<2823:ICFACM>2.0.CO;2.

——, K. W. Oleson, G. Bonan, F. Hoffman, P. Thornton, M. Vertenstein, Z. L. Yang, and X. Zeng, 2006: The Community Land Model and its climate statistics as a component of the Community Climate System Model. *J. Climate*, **19**, 2302–2324, doi:10.1175/JCLI3742.1.

Dirmeyer, P. A., X. Gao, M. Zhao, Z. Guo, T. Oki, and N. Hanasaki, 2006: GSWP-2: Multimodel analysis and implications for our perception of the land surface. *Bull. Amer. Meteor. Soc.*, **87**, 1381–1397, doi:10.1175/BAMS-87-10-1381.

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, doi:10.1029/2002JD003296.

Entin, J. K., A. Robock, K. Y. Vinnikov, S. E. Hollinger, S. Liu, and A. Namkai, 2000: Temporal and spatial scales of observed soil moisture variations in the extratropics. *J. Geophys. Res.*, **105**, 11 865–11 877, doi:10.1029/2000JD900051.

Fan, Y., H. M. van den Dool, D. Lohmann, and K. Mitchell, 2006: 1948–98 U.S. hydrological reanalysis by the Noah Land Data Assimilation System. *J. Climate*, **19**, 1214–1236, doi:10.1175/JCLI3681.1.

Flanner, M. G., C. S. Zender, J. T. Randerson, and P. J. Rasch, 2007: Present-day climate forcing and response from black carbon in snow. *J. Geophys. Res.*, **112**, D11202, doi:10.1029/2006JD008003.

Gao, F., T. He, Z. Wang, B. Ghimire, Y. Shuai, J. Masek, C. Schaaf, and C. Williams, 2014: Multiscale climatological albedo look-up maps derived from Moderate Resolution Imaging Spectroradiometer BRDF/albedo products. *J. Appl. Remote Sens.*, **8**, 083532, doi:10.1117/1.JRS.8.083532.

Gochis, D. J., W. Yu, and D. N. Yates, 2013: The WRF-Hydro model technical description and user's guide, version 1.0. NCAR Tech. Doc., 120 pp. [Available online at https://www.ral.ucar.edu/projects/wrf_hydro.]

Güntner, A., J. Stuck, S. Werth, P. Döll, K. Verzano, and B. Merz, 2007: A global analysis of temporal and spatial variations in continental water storage. *Water Resour. Res.*, **43**, W05416, doi:10.1029/2006WR005247.

Hansen, J. E., G. Russel, D. Rind, P. H. Stone, A. A. Lacis, S. Lebedeff, R. Ruedy, and L. Travis, 1983: Efficient three-dimensional dimensional global models for climate studies: Models I and II. *Mon. Wea. Rev.*, **111**, 609–662, doi:10.1175/1520-0493(1983)111<0609:ETDGMF>2.0.CO;2.

Houborg, R., M. Rodell, B. Li, R. Reichle, and B. F. Zaitchik, 2012: Drought indicators based on model-assimilated Gravity Recovery and Climate Experiment (GRACE) terrestrial water storage observations. *Water Resour. Res.*, **48**, W07525, doi:10.1029/2011WR011291.

Huang, M., Z. Hou, L. Y. R. Leung, Y. Ke, Y. Liu, Z. Fang, and Y. Sun, 2013: Uncertainty analysis of runoff simulations and parameter identifiability in the Community Land Model: Evidence from MOPEX basins. *J. Hydrometeor.*, **14**, 1754–1772, doi:10.1175/JHM-D-12-0138.1.

Jarvis, P. J., 1976: The interpretation of the variations in leaf water potential and stomatal conductance found in canopies in the field. *Philos. Trans. Roy. Soc. London*, **273B**, 593–610, doi:10.1098/rstb.1976.0035.

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.

Joyce, R. J., J. E. Janowiak, P. A. Arkin, and P. Xie, 2004: CMORPH: A method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution. *J. Hydrometeor.*, **5**, 487–503, doi:10.1175/1525-7541(2004)005<0487:CAMTPG>2.0.CO;2.

Jung, M., M. Reichstein, and A. Bondeau, 2009: Toward global empirical upscaling of FLUXNET eddy covariance observations: Validation of a model tree ensemble approach using a biosphere model. *Biogeosciences*, **6**, 2001–2013, doi:10.5194/bg-6-2001-2009.

Ke, Y., L. R. Leung, M. Huang, A. M. Coleman, H. Li, and M. S. Wigmosta, 2012: Development of high resolution land surface parameters for the Community Land Model. *Geosci. Model Dev.*, **5**, 1341–1362, doi:10.5194/gmd-5-1341-2012.

Koster, R. D., and M. J. Suarez, 1996: Energy and water balance calculations in the Mosaic LSM. NASA Tech. Memo. 104606, Vol. 9, 60 pp. [Available online at http://gmao.gsfc.nasa.gov/pubs/docs/Koster130.pdf.]

——, ——, 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, doi: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, doi:10.1016/j.envsoft.2005.07.004.

Landerer, F. W., and S. C. Swenson, 2012: Accuracy of scaled GRACE terrestrial water storage estimates. *Water Resour. Res.*, **48**, W04531, doi:10.1029/2011WR011453.

Lawrence, D. M., and Coauthors, 2011: Parameterization improvements and functional and structural advances in version 4 of the Community Land Model. *J. Adv. Model. Earth Syst.*, **3**, M03001, doi:10.1029/2011MS00045.

Leng, G., M. Huang, Q. Tang, H. Gao, and L. R. Leung, 2014: Modeling the effects of groundwater-fed irrigation on terrestrial hydrology over the conterminous United States. *J. Hydrometeor.*, **15**, 957–972, doi:10.1175/JHM-D-13-049.1.

Li, B., and M. Rodell, 2015: Evaluation of a model-based groundwater drought indicator in the conterminous U.S. *J. Hydrol.*, **526**, 78–88, doi:10.1016/j.jhydrol.2014.09.027.

——, ——, B. F. Zaitchik, R. H. Reichle, R. D. Koster, and T. M. van Dam, 2012: Assimilation of GRACE terrestrial water storage into a land surface model: EVALUATION and potential value for drought monitoring in western and central Europe. *J. Hydrol.*, **446–447**, 103–115, doi:10.1016/j.jhydrol.2012.04.035.

——, ——, and J. S. Famiglietti, 2015: Groundwater variability across temporal and spatial scales in the central and northeastern U.S. *J. Hydrol.*, **525**, 769–780, doi:10.1016/j.jhydrol.2015.04.033.

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, doi:10.1029/94JD00483.

——, Z. Xie, and M. Huang, 2003: A new parameterization for surface and groundwater interactions and its impact on water budgets with the Variable Infiltration Capacity (VIC) land surface model. *J. Geophys. Res.*, **108**, 8613, doi:10.1029/2002JD003090.

Lin, Y., and K. E. Mitchell, 2005: The NCEP stage II/IV hourly precipitation analyses: Development and applications. *19th Conf. on Hydrology*, San Diego, CA, Amer. Meteor. Soc., 1.2. [Available online at https://ams.confex.com/ams/Annual2005/techprogram/paper_83847.htm.]

Livneh, B., Y. Xia, K. E. Mitchell, M. B. Ek, and D. P. Lettenmaier, 2010: Noah LSM snow model diagnostics and enhancements. *J. Hydrometeor.*, **11**, 721–738, doi:10.1175/2009JHM1174.1.

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

Long, D., L. Longuevergne, and B. R. Scanlon, 2014: Uncertainty in evapotranspiration from land surface modeling, remote sensing, and GRACE satellites. *Water Resour. Res.*, **50**, 1131–1151, doi:10.1002/2013WR014581.

Lynch-Stieglitz, M., 1994: The development and validation of a simple snow model for the GISS GCM. *J. Climate*, **7**, 1842–1855, doi:10.1175/1520-0442(1994)007<1842:TDAVOA>2.0.CO;2.

Mesinger, F., and Coauthors, 2006: North American Regional Reanalysis. *Bull. Amer. Meteor. Soc.*, **87**, 343–360, doi:10.1175/BAMS-87-3-343.

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

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, doi:10.1029/2003JD003823.

Moody, E. G., M. D. King, C. B. Schaaf, and S. Platnick, 2008: MODIS-derived spatially complete surface albedo products: Spatial and temporal pixel distribution and zonal averages. *J. Appl. Meteor. Climatol.*, **47**, 2879–2894, doi:10.1175/2008JAMC1795.1.

Niu, G.-Y., and Z.-L. Yang, 2006: Effects of frozen soil on snowmelt runoff and soil water storage at a continental scale. *J. Hydrometeor.*, **7**, 937–952, doi:10.1175/JHM538.1.

——, ——, R. E. Dickinson, and L. E. Gulden, 2005: A simple TOPMODEL-based runoff parameterization (SIMTOP) for use in global climate models. *J. Geophys. Res.*, **110**, D21106, doi:10.1029/2005JD006111.

——, ——, ——, ——, and H. Su, 2007: Development of a Simple Groundwater Model for use in climate models and evaluation with Gravity Recovery and Climate Experiment data. *J. Geophys. Res.*, **112**, D07103, doi:10.1029/2006JD007522.

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

Oleson, K. W., and G. B. Bonan, 2000: The effects of remotely-sensed plant functional type and leaf area index on simulations of boreal forest surface fluxes by the NCAR land surface model. *J. Hydrometeor.*, **1**, 431–446, doi:10.1175/1525-7541(2000)001<0431:TEORSP>2.0.CO;2.

——, and Coauthors, 2008: Improvements to the Community Land Model and their impact on the hydrological cycle. *J. Geophys. Res.*, **113**, G01021, doi:10.1029/2007JG000563.

——, and Coauthors, 2010: Technical description of version 4.0 of the Community Land Model (CLM). NCAR Tech. Note NCAR/TN-478+STR, 257 pp., doi:10.5065/D6FB50WZ.

Otkin, J. A., and Coauthors, 2016: Assessing the evolution of soil moisture and vegetation conditions during the 2012 United States flash drought. *Agric. For. Meteor.*, **218–219**, 230–242, doi:10.1016/j.agrformet.2015.12.065.

Pelletier, J. D., and Coauthors, 2016: Global 1-km gridded thickness of soil, regolith, and sedimentary deposit layers. ORNL DAAC, accessed 13 May 2016, doi:10.3334/ORNLDAAC/1304.

Pinker, R. T., and Coauthors, 2003: Surface radiation budgets in support of the GEWEX Continental-Scale International Project (GCIP) and the GEWEX Americas Prediction Project (GAPP), including the North American Land Data Assimilation System (NLDAS) project. *J. Geophys. Res.*, **108**, 8844, doi:10.1029/2002JD003301.

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, doi:10.1175/BAMS-D-13-00263.1.

Ren, H., Z. Hou, M. Huang, J. Bao, Y. Sun, T. Tesfa, and L. R. Leung, 2016: Classification of hydrological parameter sensitivity and evaluation of parameter transferability across 431 US MOPEX basins. *J. Hydrol.*, **536**, 92–108, doi:10.1016/j.jhydrol.2016.02.042.

Reynolds, C. A., T. J. Jackson, and W. J. Rawls, 2000: Estimating soil water holding capacities by linking the Food and Agriculture Organization Soil Map of the World with global pedon databases and continuous pedotransfer functions. *Water Resour. Res.*, **36**, 3653–3662, doi:10.1029/2000WR900130.

Rienecker, M. M., and Coauthors, 2011: MERRA: NASA's Modern-Era Retrospective Analysis for Research and Applications. *J. Climate*, **24**, 3624–3648, doi:10.1175/JCLI-D-11-00015.1.

Robock, A., 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, doi:10.1175/1520-0477(2000)081<1281:TGSMDB>2.3.CO;2.

——, 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, doi:10.1029/2002JD003245.

Rodell, M., J. Chen, H. Kato, J. S. Famiglietti, J. Nigro, and C. R. Wilson, 2007: Estimating groundwater storage changes in the Mississippi River basin (USA) using GRACE. *Hydrogeol. J.*, **15**, 159–166, doi:10.1007/s10040-006-0103-7.

Rowlands, D. D., S. B. Luthcke, S. M. Klosko, F. G. R. Lemoine, D. S. Chinn, J. J. McCarthy, C. M. Cox, and O. B. Anderson, 2005: Resolving mass flux at high spatial and temporal resolution using GRACE intersatellite measurements. *Geophys. Res. Lett.*, **32**, L04310, doi:10.1029/2004GL021908.

Sakumura, C., S. Bettadpur, and S. Bruinsma, 2014: Ensemble prediction and intercomparison analysis of GRACE time-variable gravity field models. *Geophys. Res. Lett.*, **41**, 1389–1397, doi:10.1002/2013GL058632.

Sankarasubramanian, A., and R. M. Vogel, 2002: Annual hydroclimatology of the United States. *Water Resour. Res.*, **38**, 1083, doi:10.1029/2001WR000619.

Schaefer, C. L., M. H. Cosh, and T. J. Jackson, 2007: The USDA Natural Resources Conservation Service Soil Climate Analysis Network (SCAN). *J. Atmos. Oceanic Technol.*, **24**, 2073–2077, doi: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, doi: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, doi:10.1175/JTECH-D-13-00084.1.

Sellers, P. J., and Coauthors, 1996: A revised land surface parameterization (SIB2) for atmospheric GCMS. Part I: Model formulation. *J. Hydrometeor.*, **9**, 676–705, doi:10.1175/1520-0442(1996)009<0676:ARLSPF>2.0.CO;2.

Senatore, A., G. Mendicino, D. J. Gochis, W. Yu, D. N. Yates, and H. Kunstmann, 2015: Fully coupled atmosphere–hydrology simulations for the central Mediterranean: Impact of enhanced hydrological parameterization for short and long time scales. *J. Adv. Model. Earth Syst.*, **7**, 1693–1715, doi:10.1002/2015MS000510.

Slater, A. G., T. J. Bohn, J. L. McCreight, M. C. Serreze, and D. P. Lettenmaier, 2007: A multimodel simulation of pan-Arctic

hydrology. *J. Geophys. Res.*, **112**, G04S45, doi:10.1029/2006JG000303.

Stieglitz, M., A. Ducharne, R. D. Koster, and M. J. Suarez, 2001: The impact of detailed snow physics on the simulation of snow cover and subsurface thermodynamics at continental scales. *J. Hydrometeor.*, **2**, 228–242, doi:10.1175/1525-7541(2001)002<0228:TIODSP>2.0.CO;2.

Strassberg, G., B. R. Scanlon, and D. Chambers, 2009: Evaluation of groundwater storage monitoring with the GRACE satellite: Case study of the High Plains aquifer, central United States. *Water Resour. Res.*, **15**, W05410, doi:10.1029/2008WR006892.

Svoboda, M., and Coauthors, 2002: The Drought Monitor. *Bull. Amer. Meteor. Soc.*, **83**, 1181–1190, doi:10.1175/1520-0477(2002)083<1181:TDM>2.3.CO;2.

Swenson, S., and D. M. Lawrence, 2015: A GRACE-based assessment of interannual groundwater dynamics in the Community Land Model. *Water Resour. Res.*, **51**, 8817–8833, doi:10.1002/2015WR017582.

——, P. Yeh, J. Wahr, and J. Famiglietti, 2006: A comparison of terrestrial water storage variations from GRACE with in situ measurement from Illinois. *Geophys. Res. Lett.*, **33**, L16401, doi:10.1029/2006GL026962.

——, J. Famiglietti, J. Basara, and J. Wahr, 2008: Estimating profile soil moisture and groundwater variations using GRACE and Oklahoma Mesonet soil moisture data. *Water Resour. Res.*, **44**, W01413, doi:10.1029/2007WR006057.

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

Tapley, B. D., S. Bettadpur, M. Watkins, and C. Reigber, 2004: The Gravity Recovery and Climate Experiment: Mission overview and early results. *Geophys. Res. Lett.*, **31**, L09607, doi:10.1029/2004GL019920.

van Dijk, A. I. J. M., L. J. Renzullo, Y. Wada, and P. Tregoning, 2014: A global water cycle reanalysis (2003–2012) merging satellite gravimetry and altimetry observations with a hydrological multi-model ensemble. *Hydrol. Earth Syst. Sci.*, **18**, 2955–2973, doi:10.5194/hess-18-2955-2014.

Verseghy, D. L., 1991: CLASS—A Canadian land surface scheme for GCMS: I. Soil model. *Int. J. Climatol.*, **11**, 111–133, doi:10.1002/joc.3370110202.

Vuyovich, C. M., J. M. Jacobs, and S. F. Daly, 2014: Comparison of passive microwave and modeled estimates of total watershed SWE in the continental United States. *Water Resour. Res.*, **50**, 9088–9102, doi:10.1002/2013WR014734.

Wahr, J., M. Molenaar, and F. Bryan, 1998: Tine variability of the Earth's gravity field: Hydrological and oceanic effects and their possible detection using GRACE. *J. Geophys. Res.*, **103**, 30 205–30 229, doi:10.1029/98JB02844.

——, S. Swenson, V. Zlotnicki, and I. Velicogna, 2004: Time-variable gravity from GRACE: First results. *Geophys. Res. Lett.*, **31**, L11501, doi:10.1029/2004GL019779.

——, ——, and I. Velicogna, 2006: The accuracy of GRACE mass estimates. *Geophys. Res. Lett.*, **33**, L06401, doi:10.1029/2005GL025305.

Wang, A., and X. Zeng, 2009: Improving the treatment of the vertical snow burial fraction over short vegetation in the NCAR CLM3. *Adv. Atmos. Sci.*, **26**, 877–886, doi:10.1007/s00376-009-8098-3.

Xia, Y., 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, doi: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, doi:10.1029/2011JD016051.

——, 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, doi: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. Hydrometeor.*, **16**, 1962–1980, doi:10.1175/JHM-D-14-0096.1.

——, ——, ——, ——, and ——, 2015b: Comparison of NLDAS-2 simulated and NASMD observed daily soil moisture. Part II: Impact of soil texture classification and vegetation type mismatches. *J. Hydrometeor.*, **16**, 1981–2000, doi:10.1175/JHM-D-14-0097.1.

——, 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, doi:10.1175/JAMC-D-14-0275.1.

——, and Coauthors, 2016: 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, doi:10.1002/2015JD023733.

Xiao, R., X. He, Y. Zhang, V. G. Ferreira, and L. Chang, 2015: Monitoring groundwater variations from satellite gravimetry and hydrological models: A comparison with in-situ measurements in the mid-Atlantic region of the United States. *Remote Sens.*, **7**, 686–703, doi:10.3390/rs70100686.

Yang, Z.-L., and G.-Y. Niu, 2003: The Versatile Integrator of Surface and Atmosphere processes: Part 1. Model description. *Global Planet. Change*, **38**, 175–189, doi:10.1016/S0921-8181(03)00028-6.

——, R. E. Dickinson, A. Henderson-Sellers, and A. J. Pitman, 1995: Preliminary study of spin-up processes in land surface models with the first stage data of Project for Intercomparison of Land Surface Parameterization Schemes Phase 1(a). *J. Geophys. Res.*, **100**, 16 553–16 578, doi:10.1029/95JD01076.

——, and Coauthors, 2011: The community Noah land surface model with multiparameterization options (Noah-MP): 2. Evaluation over global river basins. *J. Geophys. Res.*, **116**, D12110, doi:10.1029/2010JD015140.

Zeng, X. B., and M. Decker, 2009: Improving the numerical solution of soil moisture–based Richards equation for land models with a deep or shallow water table. *J. Hydrometeor.*, **10**, 308–319, doi:10.1175/2008JHM1011.1.

Zhuo, L., D. Han, Q. Dai, T. Islam, and P. K. Srivastava, 2015: Appraisal of NLDAS-2 multi-model simulated soil moistures for hydrological modelling. *Water Resour. Manage.*, **29**, 3503–3517, doi:10.1007/s11269-015-1011-1.