# Estimates of Global Surface Hydrology and Heat Fluxes from the Community Land Model (CLM4.5) with Four Atmospheric Forcing Datasets®

AIHUI WANG

Nansen-Zhu International Research Centre, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China

# XUBIN ZENG

Department of Atmospheric Sciences, The University of Arizona, Tucson, Arizona

# DONGLIN GUO

Nansen-Zhu International Research Centre, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China

(Manuscript received 4 February 2016, in final form 19 July 2016)

#### ABSTRACT

Global land surface hydrology and heat fluxes can be estimated by running a land surface model (LSM) driven by the atmospheric forcing dataset. Previous multimodel studies focused on the impact of different LSMs on model results. Here the sensitivity of the Community Land Model, version 4.5 (CLM4.5), results to the atmospheric forcing dataset is documented. Together with the model default global forcing dataset (CRU-NCEP, hereafter CRUNCEP), three newly developed, reanalysis-based, near-surface meteorological datasets (i.e., MERRA, CFSR, and ERA-Interim) with the precipitation adjusted by the Global Precipitation Climatology Project monthly product were used to drive CLM4.5. All four simulations were run at  $0.5^{\circ} \times 0.5^{\circ}$  grids from 1979 to 2009 with the identical initialization. The simulated monthly surface hydrology variables, fluxes, and the forcing datasets were then evaluated against various observation-based datasets (soil moisture, runoff, snow depth and water equivalent, and flux tower measurements). To partially avoid the mismatch between model gridbox values and point measurements, three approaches were taken. The model simulations based on three newly constructed forcing datasets are overall better than the simulation from CRUNCEP, in particular for soil moisture and snow quantities. The ensemble mean from the CLM4.5 simulations using the four forcing datasets is generally superior to individual simulations, and the ensemble mean latent and sensible heat fluxes over global land (60°S–90°N) are 42.8 and 40.3 W m<sup>-2</sup>, respectively. The differences in both precipitation and other atmospheric forcing variables (e.g., air temperature and downward solar radiation) contribute to the differences in simulated results. The datasets are available from the authors for further evaluation and for various applications.

# 1. Introduction

As an important component of the earth system, the land surface affects the climate through its exchange of

E-mail: wangaihui@mail.iap.ac.cn

DOI: 10.1175/JHM-D-16-0041.1

water, energy, and carbon fluxes with the atmosphere. The relatively long memory of land surface processes (e.g., soil moisture) can also improve climate prediction skills (Koster et al. 2000, 2004). For example, a strong coupling between soil moisture, air temperature, and rainfall is found in monsoonal and other regions (Koster et al. 2004; Zeng et al. 2010). Furthermore, the land processes affect the partitioning of surface net radiation into latent, sensible, and ground heat fluxes (Lawrence et al. 2007). The latent heat flux (LHF) determines how much water evaporates back to the atmosphere, which influences precipitation (Makarieva et al. 2014), while the sensible heat flux (SHF) controls the air temperature, which is an important indicator of climate

Supplemental information related to this paper is available at the Journals Online website: http://dx.doi.org/10.1175/JHM-D-16-0041.s1.

*Corresponding author address*: Aihui Wang, Nansen-Zhu International Research Centre, Institute of Atmospheric Physics, Chinese Academy of Sciences, 40 Huayanli, Chaoyang District, Beijing 100029, China.

change (Ban-Weiss et al. 2011). Therefore, the global estimates of the key surface hydrological and flux variables are especially important for the land–atmosphere interaction studies.

However, in situ observations of the land surface state variables are sparse in both space and time, and some of the key variables are difficult to measure in the field. For example, in situ soil moisture data are available only at limited stations for a short period (Robock et al. 2000; Zreda et al. 2012). More recently, great efforts have been made to increase the soil moisture database, and the International Soil Moisture Network (ISMN; Dorigo et al. 2013) and the North American Soil Moisture Database (NASMD; Xia et al. 2015) are two examples. The sensible/latent heat flux measurements exist only over limited flux tower sites (https://fluxnet.ornl.gov/).

Land surface models (LSMs) are one of the useful tools used to simulate the evolution of land biogeophysical processes, and they are widely used to produce highresolution land surface state variables. The uncertainties of LSM simulations are affected by the atmospheric forcing datasets and model parameterizations. Much effort has been made to improve the model simulation through the parameterization development and incorporation of new components (e.g., Lawrence et al. 2011), and by improving the accuracy of atmospheric forcing datasets (e.g., Qian et al. 2006). However, few studies have attempted to assess the relative role of these atmospheric forcing errors and land model deficiencies (Mizukami et al. 2016).

To mitigate the uncertainties of individual model parameterization schemes, several multi-LSM projects have been developed, such as the Global Soil Wetness Project (GSWP; Dirmeyer et al. 1999, 2006), the North American Land Data Assimilation System (NLDAS; Mitchell et al. 2004; Xia et al. 2012 a,b), and the Global Land Data Assimilation System (GLDAS; Rodell et al. 2004). In these projects, several LSMs were driven by the same atmospheric forcing dataset, and the simulated land surface state variables and their ensemble mean were then evaluated (e.g., Guo et al. 2006, 2007; Xia et al. 2014). The intermodel dispersions are usually large (Dirmeyer et al. 2006), and the ensemble mean from multimodel simulations is generally superior to individual model simulations (e.g., Guo et al. 2007), because of the partial cancellation in the ensemble mean of uncertainties and biases in individual simulations. The multimodel ensemble approach has also been used in surface hydrology studies (e.g., Wang et al. 2009, 2011; van Huijgevoort et al. 2013).

While the above multimodel projects focused on model uncertainties, they did not quantify the simulation diversities caused by the uncertainties in atmospheric forcing datasets. These forcing datasets are usually derived from retrospective analyses (i.e., reanalyses) that are bias corrected by the in situ or remotely sensed observations (e.g., Qian et al. 2006; Sheffield et al. 2006). Using a revised LSM forced by the Modern-Era Retrospective Analysis for Research and Applications (MERRA) product with bias-corrected precipitation, Reichle et al. (2011) found that the land surface hydrology simulations (referred to as MERRA-Land) were greatly improved in comparison to MERRA itself. A similar study was also conducted based on European Centre for Medium-Range Weather Forecasts (ECMWF) interim reanalysis (ERA-Interim; referred to as ERAI-Land; Balsamo et al. 2015).

In this study, we have performed several sensitivity simulations using a widely used LSM driven by different global atmospheric forcing datasets to address two questions: 1) How do the variations of the atmospheric forcing dataset affect the simulated land surface state variables? and 2) Compared to observations, does the ensemble mean improve the results from individual sensitivity simulations? With a focus on the forcing dataset uncertainties, this study complements previous and ongoing community efforts that have focused on land model uncertainties (e.g., GSWP, NLDAS, and GLDAS). However, while the detailed evaluation of model results was not discussed in the same paper (i.e., these results were presented in separate papers), we decided to include the comparison of model results with comprehensive observational datasets in this paper. Section 2 introduces the model, while section 3 describes the atmospheric forcing datasets and various validation datasets. Section 4 presents the comparisons and discussion, while section 5 provides the summary.

#### 2. Land model description

The Community Land Model, version 4.5 (CLM4.5; Lawrence et al. 2011; Oleson et al. 2013), is the land component of Community Earth System Model (CESM; Hurrell et al. 2013). It simulates the terrestrial biogeophysical/ biogeochemical processes such as the water and heat transfer between vegetation and soil; the partitioning of solar and longwave radiation fluxes into latent, sensible, and ground heat fluxes; and the partitioning of precipitation into evapotranspiration, runoff, and change of total water storage. The Community Land Model with the default forcing has been extensively evaluated in numerous publications (e.g., Lawrence et al. 2011). A complete documentation of CLM4.5 is provided by Oleson et al. (2013).

The offline CLM4.5 simulations require continuous atmospheric forcing variables, including incident shortwave/longwave radiation, air temperature, humidity, pressure, precipitation, and wind speed. The atmospheric forcing dataset used in this work will be discussed in the next section. In addition, CLM4.5 needs parameters related to land use and land cover, vegetation (e.g., leaf/stem area index and fraction of vegetation coverage) and soil (e.g., soil texture). Default values of these parameters in CLM4.5 (Oleson et al. 2013) are used in this study. For simplicity, CLM4.5 is referred to as CLM hereafter.

# **3.** Atmospheric forcing datasets, validation data, and analysis methods

#### a. Atmospheric forcing datasets

Four global atmospheric forcing datasets were used here to drive CLM. One is the default dataset (CRU-NCEP, hereafter CRUNCEP) for CLM based on a combination of two existing datasets: the  $0.5^{\circ} \times 0.5^{\circ}$  monthly Climatic Research Unit (CRU) Time Series (TS; Mitchell and Jones 2005) and the  $2.5^{\circ} \times 2.5^{\circ}$  reanalysis products developed by the National Centers for Environmental Prediction and the National Center for Atmospheric Research (NCEP-NCAR) beginning in 1948 (Kalnay et al. 1996). In CRUNCEP, the air temperature, precipitation, humidity, and solar radiation have the same diurnal and day-to-day variability as the NCEP-NCAR reanalyses, but their monthly means are bias corrected by the CRU TS, version 3.1, monthly climatology. Other parameters (i.e., surface pressure, downward longwave radiation, and wind speed) are directly interpolated from NCEP-NCAR reanalyses products. Currently, there are five versions of CRUNCEP products available, derived from different versions of CRU datasets. In this work, we used version 4, which was based on CRU TS, version 3.2.1. (More details of CRUNCEP constructions are available at http://www.cesm.ucar.edu/models/cesm1.2/ clm/clm forcingdata esg.html.)

Besides CRUNCEP, three additional global atmospheric forcing datasets have been constructed based on more recent reanalysis products, including the NASA Global Modeling and Assimilation Office (GMAO) MERRA (Rienecker et al. 2011), ERA-Interim (hereafter ERAI; Dee et al. 2011), and the NCEP Climate Forecast System Reanalysis (CFSR; Saha et al. 2010). These three reanalysis products have been demonstrated to be more accurate in many aspects than the firstgeneration reanalysis (i.e., NCEP–NCAR reanalyses) because of improved models and advanced data assimilation systems (Rienecker et al. 2011).

To construct these three global forcing datasets, surface meteorology variables were downscaled to the same  $0.5^{\circ} \times 0.5^{\circ}$  grid boxes to be consistent with the spatial resolution of CRUNCEP forcing. The original grid spacing of reanalysis products is  $\frac{1}{2}^{\circ} \times \frac{1}{3}^{\circ}$  for MERRA,  $0.5^{\circ} \times 0.5^{\circ}$  for CFSR, and about  $0.75^{\circ} \times 0.75^{\circ}$  for ERAI. We used a bilinear interpolation method to downscale meteorology variables (except precipitation) to  $0.5^{\circ} \times 0.5^{\circ}$  grid boxes.

Because of the temporal heterogeneity of precipitation, bilinear interpolation would reduce the spatial variability and induce false precipitation amount. Therefore, we divided each native reanalysis grid box into  $0.01^{\circ} \times 0.01^{\circ}$ pixels with the same precipitation amount as the gridbox average at each time step. Then the precipitation at each  $0.5^{\circ} \times 0.5^{\circ}$  grid box was simply taken as the average precipitation over all pixels within this grid box. This method conserves precipitation amount within a specified area before and after mapping.

Because precipitation is the only water input in the model (without considering irrigation), errors in precipitation affect both the simulated land surface hydrological fields and the heat fluxes through evapotranspiration. Although precipitation in these three reanalysis systems is generally improved compared to earlier ones (Bosilovich et al. 2008), biases between reanalysis systems (including MERRA, ERAI, and CFSR) and the observation-based global precipitation product are still large and vary with both regions and reanalysis system (Bosilovich et al. 2011). The uncertainties in reanalysis precipitation may come from the model parameterization, the quantities and qualities of observed precipitation datasets being assimilated, and/or the assimilation schemes. Therefore, the precipitation biases in reanalysis products need to be corrected. For the same reason, only the precipitation was corrected for bias in MERRA-Land (Reichle et al. 2011) and ERAI-Land (Balsamo et al. 2015).

The Global Precipitation Climatology Project (GPCP), version 2.2,  $2.5^{\circ} \times 2.5^{\circ}$  monthly mean precipitation product (available from 1979; Adler et al. 2003; Huffman et al. 2009) was used to bias correct reanalysis precipitation in this study. Compared to the CRU precipitation based on rain gauge data only, the GPCP product was derived by merging satellite measurements and rain gauge observations. The GPCP product also uses more rain gauges than the CRU product. Trenberth et al. (2014, their Fig. 1) found that these differences affected drought reconstruction. In our bias-correction process, the GPCP product was first disaggregated to  $0.5^{\circ} \times 0.5^{\circ}$  grid boxes. Second, for each grid box, we computed a scaling factor as the ratio of monthly GPCP to monthly reanalysis precipitation. Finally, the scaling factor was multiplied by the reanalysis precipitation at each time step in the same month to form the new corrected precipitation.

It should be noted that the reanalysis precipitation was not adjusted at a small number of grid boxes with zero monthly precipitation in reanalysis but with a nonzero value in the GPCP product. This led to small differences of

	CRUNCEP	MERRA	CFSR	ERAI	Ens-mean
$Pr (mm day^{-1})$	2.08	2.33	2.33	2.34	2.27
Tair (°C)	12.27	13.88	12.96	13.14	13.31
SWd $(W m^{-2})$	196.22	203.50	194.39	190.45	196.14
$LWd(Wm^{-2})$	321.93	323.33	316.53	322.16	320.09
$ET (mm day^{-1})$	1.43	1.51	1.29	1.34	1.39
Ro $(mm day^{-1})$	0.62	0.80	1.00	0.96	0.85
SHF $(Wm^{-2})$	43.32	42.24	44.62	41.12	42.83
LHF $(W m^{-2})$	41.41	43.71	37.28	38.84	40.31
$Rn (Wm^{-2})$	85.44	87.13	83.01	81.10	84.17
$Pr - (ET + Ro) (mm day^{-1})$	-0.07	0.02	0.04	0.04	0.03

TABLE 1. Annual averages of selected quantities for 1980–2009 over global land areas (60°S–90°N).

monthly precipitation among the three new atmospheric forcing datasets (Table 1). In addition, the original temporal resolution of each reanalysis product was maintained in each forcing dataset, that is, 1-hourly for MERRA, 3-hourly for ERAI, and 6-hourly for both CFSR and CRUNCEP. During model simulation, CLM automatically interpolated each forcing variable from their input temporal resolution to the model time step (i.e., 1 h). The "coszen" (i.e., weighting based on the cosine of solar zenith angle), "nearest," and "linear" interpolation methods were used for downward shortwave radiation, precipitation, and other forcing variables, respectively (Kluzek 2013).

#### b. Datasets for land model evaluations

Instead of using measurements over one region or from one network, we used several comprehensive measurements over different land areas. These datasets include in situ soil moisture measurements in China and the United States, global river discharge, in situ snow depth and snow water equivalent (SWE) measurements in China, global reconstructed gridded snow depth data, and flux tower measurements over North and South America.

#### 1) SOIL MOISTURE

In situ measured soil moisture data are from the National Meteorological Information Center (NMIC) at the China Meteorological Administration (CMA; http:// data.cma.cn/site/index.html). The original measurements were obtained every 10 days (i.e., 8, 18, and 28 of every month) by the gravimetric technique, and no measurements were made over frozen soil. There are more than 700 stations available over China. We selected the soil moisture measurements at 10 cm below ground from 226 stations from April to September for 1993-2008, because soil moisture data below 10 cm contains too many missing values (Fig. 1). Furthermore, soil moisture measurements from Illinois, United States, were used (Hollinger and Isard 1994). This dataset contains 19 sites across the state of Illinois (37°-43°N, 88°-90°W) from February 1981 to June 2004. The soil

moisture was measured by the neutron-probe technique calibrated with gravimetric observations, and the measurements were taken at every 10 cm from the ground surface to 200 cm below the surface. In our evaluation, the monthly soil moisture measurements during 1981–2004 at three layers (i.e., 0-10, 10-100, and 100-200 cm) were used to evaluate the model simulations.

The representativeness of station soil moisture measurements in evaluating the model gridcell average is a challenging task. To partially avoid this issue, regional mean soil moisture from model simulations was compared with the site average measurements in previous studies (e.g., Xia et al. 2014). Here the same approach was applied over Illinois where the site average observed soil moisture at three depths was compared with model simulations averaged over the region (37°-43°N, 88°-90°W). Furthermore, the soil moisture sites in China were divided into three regions: northeastern China (NE; 39°-50°N, 120°-134°E), northwestern China (NW; 32°-39°N, 95°-107°E), and central eastern China (CE; 30°-39°N, 107°-120°E), where the stations are relatively dense and evenly distributed (Fig. 1). Over each region, the mean soil moisture at 0-10 cm was compared.

# 2) THE GRDC RIVER DISCHARGE

The global  $0.5^{\circ} \times 0.5^{\circ}$  monthly runoff provided by the Global Runoff Data Centre (GRDC) was used in this study (Fekete et al. 2002). The GRDC monthly runoff product is the combination of the gauge discharge data for the time period of the observation records and the outputs of a water balance model driven by observational meteorology data. The runoff is not only constrained by observed discharge values, but also preserves the spatial distribution of water balance (Fekete et al. 2002). Therefore, the data at present provide a good estimate (but not the ground truth) of global terrestrial runoff.

# 3) SNOW

Snow is one of the most important components for land surface hydrology. The model-simulated monthly



FIG. 1. The station distribution of soil moisture (226 stations), snow depth (537 stations), and SWE (342 stations) in China. The three boxes represent the three regions—northeastern China (NE; 39°–50°N, 120°–134°E), northwestern China (NW; 32°–39°N, 95°–107°E), and central eastern China (CE; 30°–39°N, 107°–120°E)—that are used in section 5b.

snow depth and SWE over China were compared to observational data provided by the NMIC (http://data. cma.cn/site/index.html). We selected 537 stations for snow depth and 342 stations for SWE from 831 stations, with the criteria that each station has data available for at least 80% of the months during 1980–2009 (Fig. 1).

#### 4) FLUX TOWER MEASUREMENTS

Decker et al. (2012) used the surface meteorological variables and fluxes at 33 FluxNet stations (https:// fluxnet.ornl.gov/) to evaluate multireanalysis products on both hourly and monthly time scales. They computed the correlation and biases of seven variables (i.e., temperature, wind speed, precipitation, downward shortwave radiation, net surface radiation, and latent and sensible heat fluxes) between reanalysis products and measurements over each station and found that the statistics vary with the comparison time scales and variables. Here the monthly LHF and SHF at 26 of the 33 stations were used (seven stations were excluded because they are all within the same  $0.5^{\circ} \times 0.5^{\circ}$  grid box). The station information is listed in Table S1 in the supplemental material. Furthermore, we used the monthly time series of selected variables at an Amazonian site (3.02°S, 53.97°W) from FluxNet.

### c. Analysis methods

We used various station observations and gridded products to evaluate model simulations. The bilinear interpolation method was used to interpolate the model grid cell to each station location. The correlation coefficient R and root-mean-square error (RMSE) were computed and compared. In section 5c, we also used the Nash–Sutcliffe efficiency coefficient  $C_{\rm NS}$  (Nash and Sutcliffe 1970), which is the ratio of mean-square error between simulated and observed values over the observed variance.

To synthesize the model performance of the four simulations over multiple stations in section 5d, we computed the probability distribution functions (PDFs) of R. To evaluate the relative performance of four simulations in sections 5c and 5e, we also used a ranking approach for  $C_{NS}$ , R, and RMSE. The ranking algorithm was used in the multidata intercomparison over the Tibetan Plateau (Wang and Zeng 2012). In the ranking algorithm,  $C_{\rm NS}$ , R, or RMSE values over each region or stations were ranked from 1 to 5 for four simulations and their ensemble mean, with 1 given to the simulation with the lowest RMSE (or highest R and  $C_{NS}$ ) and 5 with the largest RMSE (or lowest R and  $C_{NS}$ ). Then all ranking scores from all regions and all stations were averaged, and the lowest (highest) value represents the best (worst) performance of the model simulations.

## 4. Intercomparison of atmospheric forcing datasets

In the four atmospheric forcing datasets, three of them (except CRUNCEP) have nearly identical monthly precipitation as GPCP, while their diurnal intensity and frequency and day-to-day variability are still based on each reanalysis product. Figure 2 shows that the broad patterns



FIG. 2. Annual precipitation  $(mm day^{-1})$  averaged over 1980–2009 for (a) CRUNCEP, (b) GPCP, and (c) their percentage difference (GPCP minus CRUNCEP, and then divided by CRUNCEP). The mean values over global land areas are indicated. The Antarctic is excluded.

of the mean annual precipitation for the period of 1980–2009 from both GPCP and CRUNCEP generally agree with each other. However, GPCP has a global mean of  $2.33 \text{ mm day}^{-1}$ , while CRUNCEP has a smaller value of  $2.08 \text{ mm day}^{-1}$ . Over most global land areas, the annual mean difference between GPCP and CRUNCEP is positive, except over Greenland and localized areas of South America where the percentage difference is larger than 60%. The regions with the negative difference exceeding 60% in magnitude include Alaska, northern Europe, Siberia, part of eastern Asia, and South America.

Table 1 compares the simulations with different forcing datasets for a suite of variables. The mean air temperature (Tair) from 1980 to 2009 varies from 12.27°

(CRUNCEP) to 13.88°C (MERRA), incoming shortwave radiation flux (SWd) from 190.45 (ERAI) to 203.33 W m<sup>-2</sup> (MERRA), and incoming longwave radiation flux (LWd) from 316.53 (CFSR) to 323.33 W m<sup>-2</sup> (MERRA). Figures S1–S3 in the supplemental material show the spatial variability of the annual mean Tair, SWd, and LWd differences among the four products. Regions with the largest differences in magnitude for the three forcing variables are roughly consistent, located over data-sparse regions or regions with complex topography (e.g., Greenland, Tibetan Plateau, part of Africa, and South America). In contrast, the differences are relatively small over data-rich regions (e.g., North America, Europe, and eastern China), where relatively more observational data were assimilated into each reanalysis product.

The above differences in atmosphere forcing datasets affect model simulations. For example, the quality of precipitation plays an important role in the model simulations of surface hydrological variables, such as soil moisture (e.g., Wang and Zeng 2011) and surface runoff (Ro; e.g., Fekete et al. 2002). The differences in SWd also affect surface hydrology, surface temperature, and terrestrial biosphere and carbon cycle (Wild 2009; Wild and Liepert 2010). Furthermore, because CLM uses 0°C as the critical Tair to separate snowfall and rainfall, both the total precipitation (Pr) and Tair in each forcing dataset affect the snowfall (Fig. S4 in the supplemental material), which, in turn, directly affects the CLM snow simulations. Therefore, the disparities of model results for a specific process from four simulations are caused by the combined effects of all variables between different forcing datasets.

# 5. Evaluation of model simulations

In this section, the CLM-simulated global terrestrial hydrology and heat flux variables are evaluated. For model initialization, CLM was first run for 100 years, driven by cycling the CRUNCEP forcing dataset in year 1979, and the final states were saved. Then, using the above final states as the new initial fields, CLM was run from 1979 to 2009 driven by the four atmospheric forcing fields. The monthly results from 1980 to 2009 and their subsets were used in the analyses below. The ensemble mean (Ens-mean) from the four simulations was also evaluated along with individual simulations. For convenience, the outputs are referred to as CRUNCEP, MERRA, ERAI, and CFSR, respectively.

To assess the impact of different initializations on model simulations, CLM was also run for 100 years by cycling the ERAI forcing dataset for the year 1979. Indeed, the 5-yr CLM simulations driven by the ERAI forcing dataset from 1980 to 1984 with the two initial states yield very similar results (figures not shown). For instance, the difference of global mean evapotranspiration (ET) is just  $0.008 \text{ mm day}^{-1}$ , much less than the differences between simulations driven by four forcing datasets as shown in Table 1. Therefore, the disparities of different results among the four model simulations are mainly caused by the differences in forcing datasets.

# a. Annual mean quantities over global and hemispheric land

Table 1 shows that the global land annual mean runoff differs by up to  $0.38 \text{ mm day}^{-1}$ , which is greater than the maximum ET difference between four simulations

(0.22 mm day<sup>-1</sup>). The maximum LHF and net radiation (Rn) differences are comparable (6.43 and 6.03 W m<sup>-2</sup>, respectively), and both are greater than that of SHF ( $3.50 \text{ W m}^{-2}$ ). Note that  $1 \text{ mm day}^{-1}$  of ET can be roughly translated into LHF of 29 W m<sup>-2</sup>.

Since precipitation is the only water source of current model experiments (without considering irrigation), it is not surprising that both ET and Ro are different between CRUNCEP and the three other simulations. However, the identical GPCP bias-corrected precipitation in the three forcing datasets does not ensure the same variations of ET and Ro, since other forcing variables, such as Tair, are also responsible for the simulated ET and Ro (Wang and Zeng 2011; Mueller et al. 2013). Indeed the ET, SHF, and Rn differences among MERRA, CFSR, and ERAI (with the same precipitation from GPCP) are greater than those between their average and CRUNCEP. Only the Ro difference between CRUNCEP and the average of the three other simulations is greater in magnitude than differences among the three other simulations. This demonstrates the importance of both precipitation and other forcing variables on the model simulations. Furthermore, the difference of each variable in Table 1 between CRUNCEP and any of the other three simulations was found to be significant at p = 0.01 (except for ERAI Tair and LWd).

Recently, several global multiyear ET (or LHF) products based on flux tower measurements, remote sensing products, and land model outputs have been developed to facilitate the global assessments (Dirmeyer et al. 2006; Jung et al. 2010; Wang et al. 2011; Mueller et al. 2011, 2013). For example, Jung et al. (2010) estimated the global land ET of 1.57 mm day<sup>-1</sup> for 1982-2008 based on their global product constrained by flux tower measurements. This is consistent with the estimate of  $1.59 \,\mathrm{mm}\,\mathrm{day}^{-1}$  for 1989-95 (Mueller et al. 2011) based on global observations, reanalysis, and IPCC AR4 model outputs. However, based on the multidataset synthesis (including multiple global hydrological model outputs, GSWP, reanalysis, and other LSM outputs), Mueller et al. (2013) derived the global annual mean land ET of  $1.35 \text{ mm day}^{-1}$ for both 1989-95 and 1989-2005. These ET differences may be caused by the data available period, station data distribution, model biases, etc. For comparison, we also computed the annual mean ET for the same period as the above products. The Ens-mean of annual ET is  $1.39 \,\mathrm{mm}\,\mathrm{day}^{-1}$  for 1982–2008, 1.39 mm day<sup>-1</sup> for 1989–95, and  $1.40 \,\mathrm{mm}\,\mathrm{day}^{-1}$  for 1989–2005. These results agree well with that of Mueller et al. (2013), but differ from those of Jung et al. (2010) and Mueller et al. (2011).

The LHF and SHF from the four simulations (Table 1) can also be compared with those from previous modeling studies. Jiménez et al. (2011) intercompared 12 land

surface energy fluxes over global land areas for 1993–95, which include GLDAS, the second GSWP (GSWP-2), two reanalysis products, and three satellite products. Their global multiproduct means of LHF, SHF, and Rn are comparable to those in Table 1, but the spread of their multiproducts is about 20 W m<sup>-2</sup>, which is much larger than that from our four simulations. Similarly, the spread of global land mean ET for the period of 1986-95 in GSWP-2 from 13 LSMs (driven by the same atmospheric forcing dataset) is  $0.46 \,\mathrm{mm} \,\mathrm{day}^{-1}$  [based on Table 3 in Dirmeyer et al. (2006)], which is twice as large as  $0.23 \,\mathrm{mm}\,\mathrm{day}^{-1}$  from our four simulations for the same period. The larger spread of ET in GSWP-2 might be related to the larger number (i.e., 13) of models used, and the diversity of models from a simple bucket model to more comprehensive models. An alternative interpretation of the spread of results in this study and previous studies would be that surface turbulent flux differences are primarily caused by LSM differences (rather than by differences in the atmospheric forcing datasets). A coordinated international effort involving multiple state-ofthe-art land models forced by multiple atmospheric forcing datasets is needed to fully resolve this issue.

For each simulation, the ET annual anomalies were the mean of monthly ET anomalies that were computed as the global (or hemispheric) monthly ET time series minus their monthly climatology for 1980–2009. These monthly ET anomalies were then averaged to provide annual ET anomalies from all four simulations in Fig. 3. The simulations are closer to each other over the Northern Hemisphere (NH) than over the Southern Hemisphere (SH). This can be quantified by computing the standard deviation (std dev) among the four simulations in each panel in Fig. 3 in each year, and then averaging them over the whole time period. This average std dev is  $2.21 \text{ mm yr}^{-1}$  over NH, which is much smaller than 9.98 mm yr<sup>-1</sup> over SH.

The annual ET shows a positive trend in all products over the NH (all significant at p = 0.05), varying from 5.06 (MERRA) to 6.78 mm decade<sup>-1</sup> (ERAI). In contrast, the ET trends over SH are inconsistent among the four products, with negative trends in MERRA and ERAI but positive trends in CRUNCEP and CFSR. The global ET trend from Ens-mean is 7.6 mm decade<sup>-1</sup> (at p = 0.05) for 1982–97, comparable to 7.1 ± 1.0 mm decade<sup>-1</sup> from Jung et al. (2010), but the trend of -4.05 mm decade<sup>-1</sup> (at p = 0.05) for 1998–2008 is smaller in magnitude than -7.9 mm decade<sup>-1</sup> from Jung et al. (2010). The opposite sign of the trends for the two periods (1982–97 vs 1998–2007) also suggests the need for a longer period of time to draw a concrete conclusion in trend analysis.

Similar to ET, the mean annual Ro (i.e., surface plus subsurface runoff) anomalies from four simulations

(figure not shown) also show a much smaller spread over the NH (std dev =  $3.81 \text{ mm yr}^{-1}$ ) than over SH (std dev =  $11.48 \text{ mm yr}^{-1}$ ). The mean annual Ro shows positive trends in both hemispheres, with the trend in NH [varying from 1.96 for MERRA to 4.49 mm decade<sup>-1</sup> (at p = 0.05) for CRUNCEP] much smaller than over the SH [varying from 4.17 (insignificant) for ERAI to 13.3 mm decade<sup>-1</sup> (at p = 0.05) for CRUNCEP].

The spatial distribution of annual mean ET, Ro, SHF, LHF, and Rn from CRUNCEP, and the differences among four simulations are provided in Figs. S5–S9 in the supplemental material. In general, the spatial variations of these differences reflect the combined effects of forcing variables (e.g., Wang and Zeng 2011). Because the lack of global data prevents the detailed interpretation of global distribution of differences in Figs. S5–S9 in the supplemental material, here we will focus on the evaluation of four products using the datasets discussed in section 3b in the rest of this section.

# b. Soil moisture

Soil moisture is one of the most important hydrological variables on the land surface. From the surface hydrological balance equation, precipitation should be roughly balanced by the total ET and Ro over a longterm period since the change of water storage in the soil (including both soil and groundwater) is usually very small. In CLM, the surface hydrology balance also accounts for the runoff from glaciers and snow-capped surfaces (Oleson et al. 2013). Indeed, the residuals between Pr and Ro + ET in Table 1 are small (-0.07, 0.02, 0.02)0.04, and  $0.04 \,\mathrm{mm}\,\mathrm{day}^{-1}$  for CRUNCEP, MERRA, CFSR, and ERAI, respectively). Since Ro only includes surface and subsurface runoff, the small imbalance of surface hydrology is caused by the glacier and snowcapped surfaces and the slight changes of soil moisture and groundwater.

Soil moisture has a strong spatial heterogeneity partly because of the spatial heterogeneity of precipitation (e.g., Zreda et al. 2012). Therefore, the representativeness of point soil moisture measurements in evaluating model gridbox averages is always a challenging issue. To obtain robust results, here we use three approaches: 1) we use the statistics from 226 stations in China (rather than emphasize individual stations), 2) we use the average of these stations over three regions (Fig. 1), and 3) we use the average of all stations.

Table 2 shows the mean correlation coefficient of soil moisture anomalies between simulations and in situ measurements at 0–10 cm depth over 226 stations in China for the growing season (April–September) during 1993–2008. Of all simulations and their ensembles, the mean R value from CRUNCEP (R = 0.38) is smaller



FIG. 3. Global and hemisphere-averaged annual ET anomalies (mm yr<sup>-1</sup>) from model simulations. The climatology is based on the simulation from 1980 to 2009 for each atmospheric forcing dataset. The Antarctic is excluded.

TABLE 2. Mean *R* between the modeled and observed growing season (April–September) 0–10 cm soil moisture averaged for 226 stations in China from 1993 to 2008 and 19 stations in Illinois from 1981 to 2004. The correlations of snow depth and the SWE are derived for 533 stations and 342 stations in China for 1980–2009, respectively. The value in the parentheses is the percentage of stations with the correlation significant at p = 0.05.

	China			Illinois	
	Soil moisture	Snow depth	SWE	Soil moisture	
CRUNCEP	0.38 (78)	0.35 (85)	0.29 (71)	0.46 (95)	
MERRA	0.47 (89)	0.57 (95)	0.48 (87)	0.47 (95)	
CFSR	0.48 (92)	0.56 (95)	0.43 (84)	0.46 (95)	
ERAI	0.48 (91)	0.58 (97)	0.47 (88)	0.48 (95)	
Ens-mean	0.48 (90)	0.60 (97)	0.47 (88)	0.48 (95)	

than others (R = 0.47 or 0.48). For all 226 stations, the number of stations with R values significant at p = 0.05varies from 78% (CRUNCEP) to 92% (CFSR), further indicating the better performance of CFSR, ERAI, MERRA, and Ens-mean when compared with CRUNCEP.

We have also computed the R values of soil moisture anomalies between simulations and in situ measurements over three regions in China (Fig. 1). Of all simulations and their ensemble, the R values are smallest over northwestern China (e.g., 0.45 over NW for Ens-mean) than over other the two regions (e.g., 0.67 over NE and 0.64 over CE for Ens-mean). This implies that the model simulations are relatively less representative over the semiarid NW with complex topography than over the other two much more humid regions. When comparing the R values among four simulations and their ensemble mean, the R values from CRUNCEP are smallest, while Ens-mean has the overall highest R values.

Figure 4a shows the time series of monthly averaged soil moisture anomalies averaged over 226 stations in China. Except for CRUNCEP, the three other simulations are relatively consistent with each other. The four simulations overestimate the soil moisture anomalies for 1993–97, but slightly underestimate them for 2006–08. The mean *R* value of soil moisture anomalies in Fig. 4a is 0.35 for CRUNCEP, 0.46 for MERRA, and 0.47 for others, consistent with those in Table 2. Soil moisture was lowest in 1997 and highest in 1998 and 2003, reflecting the severe drought in northern China in 1997 and catastrophic floods over the Yangtze River basin in 1998 (Zong and Chen 2000).

To better understand the temporal variability of soil moisture in Fig. 4a, GPCP and CRUNCEP precipitation products were bilinearly interpolated to the 226 station locations, and the mean monthly precipitation anomalies are shown in Fig. 4b. The variations of both precipitation anomalies agree with each other very well with R of 0.64, with the largest negative value appearing in 1997 (associated with the severe drought in northern China) and positive value in 1998 (associated with the catastrophic floods over the Yangtze River basin). Soil moisture anomaly extremes in Fig. 4a lag about



FIG. 4. Monthly (a) soil moisture anomalies ( $mm^3 mm^{-3}$ ) at 0–10 cm depth and (b) precipitation anomalies ( $mm day^{-1}$ ) averaged over 226 stations in China from model simulations and observations during the growing season (April–September) for 1993–2008.



FIG. 5. Regional comparison of modeled monthly total runoff (mm day<sup>-1</sup>) averaged for 1980–2009 with GRDC climatological composite monthly runoff.

1–2 months behind precipitation anomalies in Fig. 4b. For instance, the negative extreme in precipitation anomaly appeared in June or July 1997, while the negative extreme in soil moisture anomaly was in July or August 1997. These results are consistent with many previous studies (e.g., Delworth and Manabe 1988; Wang et al. 2006).

Besides the 226 stations in China, we have also used the 19 stations in Illinois (a relatively flat area with similar climatological conditions) for 1981–2004 for model evaluations. Table 2 shows that the mean R values for soil moisture at 0–10 cm depth are close to each other (0.46–0.48) from the four simulations, and the R values at 18 out of the 19 stations are statistically significant (at p = 0.05). We also computed the R values of monthly soil moisture anomalies between simulations and in situ measurements averaged over the 19 stations in Illinois. The R values are all statistically significant (at p = 0.05) for all three layers (0–10, 10–100, and 100–200 cm). Ensmean and MERRA (e.g., with R of 0.53 for the 0–10-cm

soil layer) perform best, and CRUNCEP performs worst (with R of 0.46 for the 0–10-cm soil layer). These simulations perform much better in the 10–100-cm layer (e.g., R = 0.80 for Ens-mean) than in the other two layers (R = 0.51 or 0.53 for Ens-mean).

#### c. Runoff

We compared the simulated total runoff (surface plus subsurface runoff) climatology for 1980–2009 with GRDC composite runoff products. The GRDC global mean runoff is about 0.82 mm day<sup>-1</sup>, while our simulations vary from 0.62 (CRUNCEP) to 1.0 (CFSR) to 0.85 mm day<sup>-1</sup> (Ens-mean, Table 1). Figure 5 shows the monthly runoff averaged over 12 regions. The division of regions was similar to that in Niu et al. (2007), except that we used eastern China to replace the eastern United States. The performances of model simulations vary with both regions and atmospheric forcing datasets. For example, in western Siberia, CRUNCEP underestimated the runoff, while the three other simulations greatly overestimated runoff. All models performed better in arid regions, such as the western United States, Sahara/Arabia, and Australian arid regions.

The performances of the four simulations can also be quantified by  $C_{\rm NS}$  (as discussed in section 3c). For instance, CRUNCEP performs best over western Siberia (with  $C_{\rm NS}$  of 0.83) and worst over Congo (with  $C_{\rm NS}$ of -0.15). To evaluate the relative performance of all four simulations, we adopted a rank algorithm for  $C_{\rm NS}$  over the 12 regions in Fig. 5, as discussed in section 3c. Ens-mean has the best score (2.17), while CFSR has the worst score (3.58). When we used RMSE for the ranking, results were essentially the same, with Ens-mean having the best score (2.17) and CFSR having the worst score (3.58), demonstrating the robustness of the ranking method.

The discrepancies between modeled runoff and GRDC product may be from uncertainties in model parameterizations, the input meteorological forcing, and/or soil parameters. Decharme and Douville (2006) pointed out the importance of daily precipitation intensity in the runoff simulations. Although monthly precipitation in all four forcing datasets was corrected by observation-based datasets, the daily and diurnal variations of precipitation were the same as the original reanalysis in which the uncertainties might be large (Bosilovich et al. 2008). Over regions where runoff is mainly fed by melting snow (e.g., eastern/western Siberia, Canada, and the western United States), the snowfall and the timing of snow melting directly affect the runoff (Lawrence et al. 2012).

The snowfall/rainfall separation is determined based on Tair in CLM and hence is affected by the Tair differences among the four simulations. Therefore, CLM's simulation in runoff is related to its performance in snow processes, including snowfall, snow accumulation, and melting. Toure et al. (2016) evaluated offline CLM4simulated snow using in situ observed snow depth and satellite snow cover data and found that the model underestimated the snow depth and showed early snowmelt. We compared the SWE monthly climatology from four simulations for 1980-2009 over eastern Siberia and the western United States (figure not shown) and found that SWE peaked in April and February, respectively, which were much earlier than in the observations (e.g., Lopez Caceres et al. 2015). This explains the model runoff peak in May (vs observed peak in June) over eastern Siberia and the model runoff peak in March (vs observed peak in May) over the western United States in Fig. 5.

Moreover, the inconsistency of time periods between our simulations and the GRDC product, and the density of observed runoff data used in GRDC, are also responsible for these differences. Since the observed river discharge is affected by human activities (e.g., dams and agricultural, industrial, and household consumptions of water), the GRDC dataset may also underestimate the natural annual mean runoff and affect the seasonal cycle of monthly runoff in Fig. 5. Furthermore, the uncertainties in the surface meteorological forcing dataset and the use of a simple water balance model may also affect the accuracy of composite GRDC product.

#### d. Snow depth and SWE

For the snow evaluation, the observed monthly snow depth (537 stations) and SWE (342 stations) in China during major snow seasons (October-March) for 1980-2009 were used to evaluate model simulations. Table 2 shows that the mean R values averaged at all stations are the worst from CRUNCEP (0.35 for snow depth and 0.29 for SWE), while results from the other three simulations are relatively close to each other. Ens-mean gives the overall best results (0.60 for snow depth and 0.47 for SWE). The R values over most stations are statistically significant (at p = 0.05), and the percentage of station numbers with statistically significant R in CRUNCEP (85% for snow depth and 71% for SWE) is less than those from other simulations (more than 95%) for snow depth and more than 84% for SWE). For each simulation, Table 2 also shows that its performance on snow depth is better than SWE.

We have also compared these performances with that of the Canadian Meteorological Centre (CMC) snow depth product (Brasnett 1999; Brown and Brasnett 2010) using the in situ measurements in China. CMC has a mean R of 0.52, which is higher than that of CRUNCEP but lower than other simulations in Table 2. CMC also shows a statistically significant R (at p = 0.05) at 78% of stations only, which is lower than all four simulations in Table 2.

To further interpret these results from hundreds of stations, Fig. 6 plots the cumulative distribution function (CDF) of R for snow depth and SWE. For both panels, the R values from CRUNCEP are smallest among all simulations, while the performances of other simulations are close to each other. Ens-mean is overall better than most of the simulations. For example, the percentage of stations with R above 0.5 of snow depth is about 67% for Ens-mean, while this number is 61% for MERRA, 62% for both CFSR and ERAI, and 19% only for CRUNCEP. Similarly, the percentage of stations with R above 0.5 of SWE is 46% for Ens-mean, while this percentage is 47% for MERRA, 38% for CFSR, 45% for ERAI, and just 11% for CRUNCEP. These results are consistent with those in Table 2.

The deficiencies of model snow process parameterization and biases in the meteorology forcing dataset are two major factors that affect the snow modeling. CLM uses  $0^{\circ}$ C as the critical Tair to separate snowfall and



FIG. 6. CDF of the correlation between station observations and model simulations from 1980 to 2009 for (a) monthly snow depth (using 576 stations) and (b) monthly SWE (using 342 stations) in China.

rainfall. In the four simulations, we computed the mean snowfall averaged over all 537 stations with SWE measurements and found that both the snowfall amount and monthly variability in CRUNCEP were much smaller than others (figure not shown), which is related to its poor performance in Fig. 6. Moreover, the inconstancies between models and observations might also be responsible for the better performance on snow depth than SWE in all four simulations. In CLM, whenever modeled snow is present, SWE is greater than zero. However, in the observational data, when the observed snow depth is less than 4mm, SWE is set to be zero. Furthermore, snow depth in CLM is a diagnostic variable, computed from the SWE and snow density, which is dependent on Tair and snow aging. In contrast, the observed snow depth is directly measured, while SWE is derived from snow depth and the lookup table snow density.

#### e. Flux tower measurements

The representativeness of flux tower measurements in evaluating model gridbox variables is also a challenging issue. To increase the robustness of our results, we

emphasize the statistics based on the 26 AmeriFlux sites (Table S1 in the supplemental material). Table 3 lists the mean R and RMSE of LHF and SHF and their ranking scores. For SHF, the mean R varies from 0.70 (CFSR) to 0.76 (ERAI), and the mean RMSE varies from 20.85 (CRUNCEP) to  $26.85 \text{ W m}^{-2}$  (CFSR). For LHF, the mean R is similar (0.81–0.83), and the mean RMSE is also similar  $(19.58-21.45 \text{ W m}^{-2})$  among the simulations. These values are just slightly worse than those (R = 0.84;  $RMSE = 17.49 W m^{-2}$ ) between the in situ data and the Jung et al. (2010) product that was constrained by flux tower measurements (including the 26 AmeriFlux sites used here). From the ranking scores, the best SHF score is ERAI for R and Ens-mean for RMSE, while the worst one is CFSR for both R and RMSE. The best LHF score is Ens-mean for both R and RMSE, while the worst score is CRUNCEP for R and MERRA for RMSE. Based on the average of all four scores (Table 3), Ensmean performs best (with a mean score of 2.19) and CRUNCEP performs worst (with a mean score of 3.40).

A detailed comparison with flux tower in situ measurements was also conducted at the K83 rain forest site

TABLE 3. Mean R and RMSE (W m<sup>-2</sup>) between the modeled and measured LHF and SHF over 26 AmeriFlux tower sites. The scores are explained in section 3c, and the smaller the score, the better.

	SHF			LHF				
	Mean R	Mean RMSE	R score	RMSE score	Mean R	Mean RMSE	R score	RMSE score
CRUNCEP	0.72	20.85	3.73	2.69	0.81	20.82	3.92	3.27
MERRA	0.74	23.65	2.92	3.00	0.82	21.45	3.00	4.00
CFSR	0.70	26.85	3.85	4.15	0.81	19.58	3.00	2.42
ERAI	0.76	23.58	2.12	3.04	0.82	20.48	3.08	3.04
Ens-mean	0.75	22.25	2.38	2.12	0.83	19.66	2.00	2.27



FIG. 7. Comparisons of monthly mean quantities among model forcing/simulations and observations at an Amazonian site  $(3.02^{\circ}S, 53.97^{\circ}W)$  from July 2000 to December 2003. The variables include (a) precipitation (mm), (b) air temperature (°C), (c) downward shortwave radiation (W m<sup>-2</sup>), (d) downward longwave radiation (W m<sup>-2</sup>), (e) SHF (W m<sup>-2</sup>), and (f) LHF (W m<sup>-2</sup>) in which the results from Jung et al. (2010) are also included.

over the Amazon (3.02°S, 53.97°W). Over the relatively homogeneous tropical rain forest, this site complements the AmeriFlux sites in Table S1 in the supplemental material. There is a strong precipitation seasonal cycle, with the peak in austral summer (Fig. 7a). The average precipitation is 314 mm month<sup>-1</sup> from observation for the study period, while both GPCP and CRUNCEP are close to observation with about  $32 \text{ mm month}^{-1}$  overestimation. Figure 7b shows that all forcing datasets overestimated monthly Tair, and CFSR products were closest to the observations with a mean bias of 0.51°C. The seasonal cycle of Tair from ERAI was much greater than the observational variation. The monthly SWd from the four products showed different seasonal variation from each other and from observations (Fig. 7c). Averaged for the whole period, the mean bias of four products in SWd varies from -5.40 (CFSR) to  $38.98 \text{ W} \text{ m}^{-2}$  (ERAI). While ERAI monthly LWd agreed very well with observations, the three other products overestimated it (Fig. 7d).

While the observed monthly SHF was around  $20 \text{ W m}^{-2}$  with small seasonal variations, all products showed stronger seasonal variations and substantially overestimated SHF (Fig. 7e). This overestimate may be related to the overestimate of SWd and LWd in Figs. 7c and 7d. In contrast to the large SHF overestimation, some of the products showed very small mean LHF bias

(e.g.,  $-1.8 \text{ Wm}^{-2}$  for ERAI and  $2.8 \text{ Wm}^{-2}$  for CRUNCEP; Fig. 7f). It is interesting to note that, although the K83 observations were used in the Jung et al. (2010) product, the mean LHF bias from their product ( $-4.0 \text{ Wm}^{-2}$ ) is still larger in magnitude than that from ERAI and CRUNCEP (Fig. 7f). Similarly, RMSE of LHF from Jung et al. (2010) (11.9 Wm<sup>-2</sup>) is larger than that from CRUNCEP (9.9 Wm<sup>-2</sup>) and CFSR (11.5 Wm<sup>-2</sup>) products. As discussed in Jung et al. (2010), their LHF biases may be from the construction method, the input meteorology dataset, and other factors.

#### 6. Conclusions

Reliable land surface hydrological variables simulated from LSMs depend on accurate atmospheric forcing datasets. Because of a lack of high-resolution globally observed meteorological variables, we have to rely on the reanalysis products to construct the global LSM forcing datasets. In this paper, CLM was driven by three reanalysis-based surface meteorological datasets (i.e., MERRA, CFSR, and ERAI), as well as the premerged reanalysis with CRU product (CRUNCEP). Following MERRA-Land and ERAI-Land in which only precipitation was bias corrected, the three reanalysis precipitation datasets were adjusted by the GPCP monthly product because of their known deficiencies, and then all simulations were conducted at the same horizontal resolution  $0.5^{\circ} \times 0.5^{\circ}$  with the same initialization. The modelsimulated monthly surface hydrology variables (e.g., snow depth, SWE, runoff, and soil moisture), fluxes (e.g., latent and sensible heat fluxes), and their ensemble mean were then intercompared and evaluated against various observation-based datasets (soil moisture, runoff, snow depth and SWE, and flux tower measurements).

The model simulations based on three newly constructed forcing datasets perform better than the simulation from CRUNCEP (the default forcing dataset in CLM is partly due to the data availability for a longer period of time), in particular for soil moisture and snow quantities. For example, the mean correlation coefficient between model-simulated soil moisture and observations over 226 stations in China is only 0.38 for CRUNCEP, but 0.47-0.48 for others. This is partially due to the use of meteorological data from three new reanalyses (MERRA, CFSR, and ERAI), which are more accurate than the first-generation reanalysis product (i.e., NCEP-NCAR reanalyses). This also implies that the realistic atmospheric forcing dataset is important for model simulations. Note that this conclusion is based on CLM, and more studies using other LSMs are needed to test its robustness. Furthermore, the CRUNCEP dataset remains useful as it covers a longer period (from 1948) than the other forcing datasets (from 1979). The CRUNCEP is also available from 1901 to 1947, but the NCEP-NCAR reanalyses product for the same year (1948) was used for the adjustment for each year (i.e., interannual variability was neglected).

Although precipitation in the three new forcing datasets was adjusted by the same GPCP monthly product, the global mean differences of hydrological variables are not always smaller than their differences from CRUNCEP in magnitude. For instance, the maximum ET difference among the three new simulations is 0.22 mm day<sup>-1</sup>, while the difference is within 0.14 mm day<sup>-1</sup> between CRUNCEP and others. This indicates that the partitioning of precipitation between runoff and evapotranspiration is affected by the combination of all forcing variables.

Because the ensemble mean of multimodel simulations is usually shown to be closer to observations than individual model simulations, it is widely used in data developments, drought reconstruction, and climate change research (e.g., Guo et al. 2007; Wang et al. 2009, 2011; Mueller et al. 2013). In this study, the ensemble mean is also found to be overall superior to the individual forcing simulations, consistent with findings in previous studies. For example, the mean *R* derived from snow depth in China is 0.60 for Ens-mean, which is larger than those from individual simulations, varying from 0.35 (CRUNCEP) to 0.58 (ERAI). Similar performances were also found in both SWE and soil moisture simulations. The global runoff from Ens-mean  $(0.85 \text{ mm day}^{-1})$  is also closer to GRDC composite dataset  $(0.82 \text{ mm day}^{-1})$  than those from individual simulations. These results further demonstrate the value of using ensemble mean in the representations of land surface hydrology and heat fluxes.

Based on the ensemble mean, the energy and water balance over global land  $(60^{\circ}\text{S}-90^{\circ}\text{N})$  is precipitation  $(2.27 \text{ mm day}^{-1})$  nearly balanced by ET  $(1.39 \text{ mm day}^{-1})$ and Ro  $(0.85 \text{ mm day}^{-1})$ , and net radiative flux  $(84.17 \text{ Wm}^{-2})$  nearly balanced by SHF  $(42.83 \text{ Wm}^{-2})$ and LHF  $(40.31 \text{ Wm}^{-2})$ . The above ET estimate is consistent with the recent estimate of Mueller et al. (2013)  $(1.35 \text{ mm day}^{-1})$  but smaller than earlier estimates  $[1.59 \text{ mm day}^{-1}]$  from Mueller et al. (2011) and  $1.57 \text{ mm day}^{-1}$  from Jung et al. (2010)]. Our estimates of the annual mean SHF, LHF, and Rn from Ens-mean are also consistent with the multiproduct means from Jiménez et al. (2011).

The accuracy of the offline LSM simulations depends on both model parameterization schemes (including input parameters) and the atmospheric forcing datasets. Previous projects on offline LSMs, such as GLDAS (Rodell et al. 2004) and GSWP (Dirmeyer et al. 1999, 2006), were all based on the same forcing dataset to drive multiple LSMs. They focused on simulations due to the model parameterization discrepancy, but simulations due to uncertainties in the atmospheric forcing dataset were less concerned (Guo et al. 2006). Complementary to these efforts, this study emphasized the impact of uncertainties in forcing datasets on the CLM simulations. It is found that the spread of the global mean ET and Ro from the four simulations is much smaller than that from 13 models in GSWP-2. While this could be caused by the much smaller number of forcing datasets used in this study, it could also imply that the diversity in model parameterizations (rather than in atmospheric forcing dataset) has a dominant effect on the dispersion of model results, and efforts should be made to improve the LSM parameterizations. Community efforts (like GSWP and GLDAS) are needed in the future to run several land models driven by different atmospheric forcing datasets to separate the impacts of land models and forcing dataset.

Besides model deficiencies and uncertainties in atmospheric forcing dataset, the mismatch between model gridbox values and point measurements is another reason for the differences between model results and observationbased products. To partially avoid this issue and obtain robust results, we have taken three approaches in model evaluations: using the statistics from point measurements, using the spatial average of point measurements over specific regions, and combining point measurements with observation-based gridded datasets.

In this paper, we have produced the global  $0.5^{\circ} \times 0.5^{\circ}$ land surface hydrology, fluxes, and other land surface quantities from 1980 to 2009. Given the uncertainties in both LSMs and atmospheric forcing datasets, our datasets are complementary to those produced by different LSMs using the same atmospheric forcing dataset (e.g., from GLDAS and GSWP). Therefore, our datasets from 1979 to 2009 (available from the authors), including both the four forcing datasets and their simulations, can be combined with those from GLDAS and GSWP for a variety of applications, for example, for the study of global or regional hydrological cycle, soil moisture memory, and soil moisture–groundwater interactions.

Acknowledgments. The work of A.W. was supported by the National Science Foundation of China (NSFC) under Grant 41275110, the work of X.Z. was supported by the NSF (AGS-0944101), and the work of D.G. was support by the NSFC under Grant 41405087. The authors thank Dr. Youlong Xia and two anonymous reviewers for their constructive comments. We thank the data centers for providing various datasets for this work. The ERA-Interim products were obtained from the NCAR Research Data Archive. The MERRA products were from the Goddard Earth Sciences (GES) Data and Information Services Center (DISC). The CFSR products were obtained from NOAA's National Operational Mode Archive and Distribution System (NOMADS), which is maintained at NOAA's National Climatic Data Center (NCDC). The CRUNCEP dataset was from http://dods.extra.cea.fr/data/p529viov/cruncep/. The soil moisture and snow data in China were from http://data. cma.cn/site/index.html.

#### REFERENCES

- Adler, R. F., and Coauthors, 2003: The version-2 Global Precipitation Climatology Project (GPCP) monthly precipitation analysis (1979–present). J. Hydrometeor., 4, 1147–1167, doi:10.1175/1525-7541(2003)004<1147:TVGPCP>2.0.CO;2.
- Balsamo, G., and Coauthors, 2015: ERA-Interim/Land: A global land surface reanalysis data set. *Hydrol. Earth Syst. Sci.*, 19, 389–407, doi:10.5194/hess-19-389-2015.
- Ban-Weiss, G. A., G. Bala, L. Cao, J. Pongratz, and K. Caldeira, 2011: Climate forcing and response to idealized changes in surface latent and sensible heat. *Environ. Res. Lett.*, 6, 034032, doi:10.1088/1748-9326/6/3/034032.
- Bosilovich, M. G., J. Chen, F. R. Robertson, and R. F. Adler, 2008: Evaluation of global precipitation in reanalyses. *J. Appl. Meteor. Climatol*, **47**, 2279–2299, doi:10.1175/ 2008JAMC1921.1.
- —, F. R. Robertson, and J. Chen, 2011: Global energy and water budgets in MERRA. J. Climate, 24, 5721–5739, doi:10.1175/ 2011JCLI4175.1.

- Brasnett, B., 1999: A global analysis of snow depth for numerical weather prediction. *J. Appl. Meteor.*, **38**, 726–740, doi:10.1175/ 1520-0450(1999)038<0726:AGAOSD>2.0.CO;2.
- Brown, R. D., and B. Brasnett, 2010: Canadian Meteorological Centre (CMC) Daily Snow Depth Analysis Data. National Snow and Ice Data Center, accessed 17 August 2016. [Available online at http://nsidc.org/data/nsidc-0447.html.].
- Decharme, B., and H. Douville, 2006: Uncertainties in the GSWP-2 precipitation forcing and their impacts on regional and global hydrological simulations. *Climate Dyn.*, 27, 695–713, doi:10.1007/s00382-006-0160-6.
- Decker, M., M. A. Brunke, Z. Wang, K. Sakaguchi, X. Zeng, and M. G. Bosilovich, 2012: Evaluation of the reanalysis products from GSFC, NCEP, and ECMWF using flux tower observations. J. Climate, 25, 1916–1944, doi:10.1175/ JCLI-D-11-00004.1.
- Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, doi:10.1002/qj.828.
- Delworth, T. L., and S. Manabe, 1988: The influence of potential evaporation on the variabilities of simulated soil wetness and climate. J. Climate, 1, 523–547, doi:10.1175/ 1520-0442(1988)001<0523:TIOPEO>2.0.CO;2.
- Dirmeyer, P. A., A. J. Dolman, and N. Sato, 1999: The Global Soil Wetness Project: A pilot project for global land surface modeling and validation. *Bull. Amer. Meteor. Soc.*, **80**, 851–878, doi:10.1175/1520-0477(1999)080<0851:TPPOTG>2.0.CO;2.
- —, 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.
- Dorigo, W. A., and Coauthors, 2013: Global automated quality control of in situ soil moisture data from the International Soil Moisture Network. *Vadose Zone J.*, 12, doi:10.2136/vzj2012.0097.
- Fekete, B. M., C. J. Vorosmarty, and W. Grabs, 2002: Highresolution fields of global runoff combining observed river discharge and simulated water balances. *Global Biogeochem. Cycles*, **16**, 1042, doi:10.1029/1999GB001254.
- Guo, Z., P. A. Dirmeyer, Z. Hu, X. Gao, and M. Zhao, 2006: Evaluation of the second Global Soil Wetness Project soil moisture simulations. 2: Sensitivity to external meteorological forcing. *J. Geophys. Res.*, **111**, D22S03, doi:10.1029/2006JD007845.
- —, —, X. Gao, and M. Zhao, 2007: Improving the quality of simulated soil moisture with a multi-model ensemble approach. *Quart. J. Roy. Meteor. Soc.*, **133**, 731–747, doi:10.1002/qj.48.
- Hollinger, S. E., and S. A. Isard, 1994: A soil moisture climatology of Illinois. J. Climate, 7, 822–833, doi:10.1175/ 1520-0442(1994)007<0822:ASMCOI>2.0.CO;2.
- Huffman, G. J., R. F. Adler, D. T. Bolvin, and G. Gu, 2009: Improving the global precipitation record: GPCP version 2.1. *Geophys. Res. Lett.*, 36, L17808, doi:10.1029/2009GL040000.
- Hurrell, J., and Coauthors, 2013: The Community Earth System Model: A framework for collaborative research. *Bull. Amer. Meteor. Soc.*, 94, 1339–1360, doi:10.1175/BAMS-D-12-00121.1.
- Jiménez, C., and Coauthors, 2011: Global intercomparison of 12 land surface heat flux estimates. J. Geophys. Res., 116, D02102, doi:10.1029/2010JD014545.
- Jung, M., and Coauthors, 2010: Recent decline in the global land evapotranspiration trend due to limited moisture supply. *Nature*, 467, 951–954, doi:10.1038/nature09396.
- Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. Bull. Amer. Meteor. Soc., 77, 437–471, doi:10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.

- Kluzek, E., 2013: CESM research tools: CLM4.5 in CESM1.2.0 user's guide documentation. NCAR, accessed 17 August 2016. [Available online at http://www.cesm.ucar.edu/models/cesm1.2/ clm/models/lnd/clm/doc/UsersGuide/book1.html.]
- Koster, R. D., M. J. Suarez, and M. Heiser, 2000: Variance and predictability of precipitation at seasonal-to-interannual timescales. *J. Hydrometeor.*, **1**, 26–46, doi:10.1175/1525-7541(2000)001<0026: VAPOPA>2.0.CO;2.
- —, and Coauthors, 2004: Regions of strong coupling between soil moisture and precipitation. *Science*, **305**, 1138–1140, doi:10.1126/ science.1100217.
- Lawrence, D., P. E. Thornton, K. W. Oleson, and G. B. Bonan, 2007: Partitioning of evaporation into transpiration, soil evaporation, and canopy evaporation in a GCM: Impacts on land-atmosphere interaction. J. Hydrometeor., 8, 862–880, doi:10.1175/JHM596.1.
- —, 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:10.1029/2011MS000045.
- —, K. W. Oleson, M. G. Flanner, C. G. Fletcher, P. J. Lawrence, S. Levis, S. C. Swenson, and G. B. Bonan, 2012: The CCSM4 land simulation, 1850–2005: Assessment of surface climate and new capabilities. *J. Climate*, **25**, 2240–2260, doi:10.1175/ JCLI-D-11-00103.
- Lopez Caceres, M., F. Takakai, G. Iwahana, A. N. Fedorov, Y. Iijima, R. Hatano, and M. Fukuda, 2015: Snowmelt and the hydrological interaction of forest–grassland ecosystems in central Yakutia, eastern Siberia. *Hydrol. Processes*, **29**, 3074–3083, doi:10.1002/hyp.10424.
- Makarieva, A. M., V. G. Gorshkov, D. Sheil, A. D. Nobre, P. Bunyard, and B.-L. Li, 2014: Why does air passage over forest yield more rain? Examining the coupling between rainfall, pressure, and atmospheric moisture content. J. Hydrometeor., 15, 411–426, doi:10.1175/JHM-D-12-0190.1.
- Mitchell, J. D., and P. D. Jones, 2005: An improved method of constructing a database of monthly climate observations and associated high-resolution grids. *Int. J. Climatol.*, 25, 693–712, doi:10.1002/joc.1181.
- 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.
- Mizukami, N., and Coauthors, 2016: Implications of the methodological choices for hydrologic portrayals of climate change over the contiguous United States: Statistically downscaled forcing data and hydrologic models. J. Hydrometeor., 17, 73–98, doi:10.1175/JHM-D-14-0187.1.
- Mueller, B., and Coauthors, 2011: Evaluation of global observationsbased evapotranspiration datasets and IPCC AR4 simulations. *Geophys. Res. Lett.*, 38, L06402, doi:10.1029/2010GL046230.
- —, and Coauthors, 2013: Benchmark products for land evapotranspiration: LandFlux-EVAL multi-dataset synthesis. *Hydrol. Earth Syst. Sci.*, 17, 3707–3720, doi:10.5194/hess-17-3707-2013.
- Nash, J. E., and J. V. Sutcliffe, 1970: River flow forecasting through conceptual models. Part I—A discussion of principles. *J. Hydrol.*, **10**, 282–290, doi:10.1016/0022-1694(70)90255-6.
- Niu, G.-Y., Z.-L. Yang, R. E. Dickinson, L. E. Gulden, 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.

- Oleson, K., and Coauthors, 2013: Technical description of version 4.5 of the Community Land Model (CLM). NCAR Tech. Note NCAR/TN-503+STR, 420 pp., doi:10.5065/D6RR1W7M.
- Qian, T., A. Dai, K. E. Trenberth, and K. W. Oleson, 2006: Simulation of global land surface conditions from 1948 to 2004: Part I: Forcing data and evaluations. *J. Hydrometeor.*, 7, 953–975, doi:10.1175/JHM540.1.
- Reichle, R., R. D. Koster, J. Gabriëlle, G. De Lannoy, B. Forman, Q. Liu, S. P. P. Mahanama, and A. Touré, 2011: Assessment and enhancement of MERRA land surface hydrology estimates. *J. Climate*, 24, 6322–6338, doi:10.1175/JCLI-D-10-05033.1.
- Rienecker, M. R., 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. Vinnikov, G. Srinivasan, J. Entin, S. Hollinger, N. 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.
- Rodell, M., and Coauthors, 2004: The Global Land Data Assimilation System. Bull. Amer. Meteor. Soc., 85, 381–394, doi:10.1175/ BAMS-85-3-381.
- Saha, S., and Coauthors, 2010: The NCEP Climate Forecast System Reanalysis. Bull. Amer. Meteor. Soc., 91, 1015–1057, doi:10.1175/ 2010BAMS3001.1.
- Sheffield, J., G. Goteti, and E. F. Wood, 2006: Development of a 50-year high-resolution global dataset of meteorological forcings for land surface modeling. *J. Climate*, **19**, 3088–3111, doi:10.1175/JCLI3790.1.
- Toure, A., M. Rodell, Z. Yang, H. Beaudoing, E. Kim, Y. Zhang, and Y. Kwon, 2016: Evaluation of the snow simulations from the Community Land Model, version 4 (CLM4). *J. Hydrometeor.*, 17, 153–170, doi:10.1175/JHM-D-14-0165.1.
- Trenberth, K. E., A. Dai, G. van der Schrier, P. D. Jones, J. Barichivich, K. R. Briffa, and J. Sheffield, 2014: Global warming and changes in drought. *Nat. Climate Change*, 4, 17–22, doi:10.1038/nclimate2067.
- van Huijgevoort, M. H. J., and Coauthors, 2013: Global multimodel analysis of drought in runoff for the second half of the twentieth century. J. Hydrometeor., 14, 1535–1552, doi:10.1175/ JHM-D-12-0186.1.
- Wang, A., and X. Zeng, 2011: Sensitivities of terrestrial water cycle simulations to the variations of precipitation and air temperature in China. J. Geophys. Res., 116, D02107, doi:10.1029/ 2010JD014659.
- —, and —, 2012: Evaluation of multi reanalysis products with in situ observations over the Tibetan Plateau. J. Geophys. Res., 117, D05102, doi:10.1029/2011JD016553.
- —, —, S. S. P. Shen, Q.-C. Zeng, and R. Dickinson, 2006: Timescales of land surface hydrology. J. Hydrometeor., 7, 868–879, doi:10.1175/JHM527.1.
- —, T. Bohn, P. Mahanama, R. Koster, and D. Lettenmaier, 2009: Multimodel ensemble reconstruction of drought over the continental United States. *J. Climate*, **22**, 2694–2712, doi:10.1175/ 2008JCLI2586.1.
- —, D. Lettenmaier, and J. Sheffield, 2011: Soil moisture drought in China, 1950–2006. J. Climate, 24, 3257–3270, doi:10.1175/ 2011JCLI3733.1.
- Wild, M., 2009: Global dimming and brightening: A review. J. Geophys. Res., 114, D00D16, doi:10.1029/2008JD011470.
- —, and B. Liepert, 2010: The earth radiation balance as driver of the global hydrological cycle. *Environ. Res. Lett.*, 5, 025203, doi:10.1088/1748-9326/5/2/025203.

- 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. Ek, J. Dong, N. Chaney, H. Wei, J. Meng, and E. 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.
- —, T. W. Ford, Y. Wu, S. M. Quiring, and M. B. Ek, 2015: 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.
- Zeng, X., M. Barlage, C. Castro, and K. Fling, 2010: Comparison of land-precipitation coupling strength using observations and models. J. Hydrometeor., 11, 979–994, doi:10.1175/ 2010JHM1226.1.
- Zong, Y., and X. Chen, 2000: The 1998 flood on the Yangtze, China. *Nat. Hazards*, **22**, 165–184, doi:10.1023/A:1008119805106.
- Zreda, M., W. J. Shuttleworth, X. Zeng, C. Zweck, D. Desilets, T. Franz, and R. Rosolem, 2012: COSMOS: Cosmic-Ray Soil Moisture Observing System. *Hydrol. Earth Syst. Sci.*, **16**, 4079–4099, doi:10.5194/hess-16-4079-2012.