Journal of Hydrology 595 (2021) 126021

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Understanding the spatial patterns of evapotranspiration estimates from land surface models over China

Ruochen Sun [a,b], Qingyun Duan [a,b,*], Jiahu Wang [b]

[a] State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Hohai University, Nanjing, China
[b] College of Hydrology and Water Resources, Hohai University, Nanjing, China

ARTICLE INFO

This manuscript was handled by Emmanouil Anagnostou, Editor-in-Chief

Keywords:
Evapotranspiration
Land surface model
GLDAS
Spatial evaluation
China

ABSTRACT

Land surface models are important tools to represent and predict the spatiotemporal variability of evapotranspiration, which is a key variable in terrestrial water, energy and carbon cycles. However, evapotranspiration estimates from land surface models may suffer from various uncertainties in land surface modeling. Therefore, assessing the performance of evapotranspiration simulation plays a vital role in understanding the deficiencies of land surface modeling. Most of the evaluation studies of modeled evapotranspiration relied on comparisons with flux site observations (point scale) and water budget-derived evapotranspiration (basin scale), which have certain drawbacks and limitations. Thus, the evaluation results may be misleading for understanding the performance of land surface models on representing the spatial variability of evapotranspiration. In this study, a thorough spatial evaluation of the new and reprocessed Global Land Data Assimilation System evapotranspiration products is performed across China based on three bias-insensitive spatial evaluation methods, including the empirical orthogonal function analysis, the connectivity analysis and the fractions skill score. These evapotranspiration products were estimated from three land surface models, namely Noah, VIC and CLSM. The conventional evapotranspiration evaluation against eddy covariance measurements is also performed. The results show that all three products have consistent trends with the observed evapotranspiration series at both daily and monthly time scales. Noah and VIC have comparable performances in terms of different statistic metrics and outperform CLSM at both time scales. The spatial evaluation methods can provide additional valuable information to diagnose the model errors. VIC has the worst spatial performance during the warm months. Despite its inferior performance in late winter and early spring, Noah, overall, has the best spatial performance among the three. The gained insights of this study can help to improve the spatial performance of these models and further promote the system development.

1. Introduction

Terrestrial evapotranspiration (ET), which consists of evaporation from soil and canopy interception, vegetation transpiration and sublimation of ice and snow, consumes two-thirds of total global land surface precipitation (Oki and Kanae, 2006). ET plays a critical role in the exchange of water, energy and carbon among hydrosphere, atmosphere, pedosphere and biosphere (Fisher, 2014; Katul et al., 2012; Wang and Dickinson, 2012). It is also the key variable in linking the ecological and hydrological process (Fisher et al., 2017). Therefore, reliable and accurate ET estimates are vital for understanding the impact on local weather (Miralles et al., 2014), monitoring extreme event droughts (Anderson et al., 2013; Vicente-Serrano et al., 2018), diagnosing climate variability and change (Deb et al., 2019; Mao et al., 2015; Sheffield et al., 2012), improving water resource management (Anderson et al., 2011, 2012) and so on.

Currently, there have been various methods developed to estimate ET at point or field (from regional to global) scales. These methods can be grouped into five main categories: (1) the water balance method (Liu et al., 2016; Zeng et al., 2014), (2) upscaling of eddy covariance (EC) flux measurements at tower sites (Jung et al., 2011; Li et al., 2018; Xu et al., 2018), (3) machine learning-based ET estimation (Adnan et al., 2020; Alizamir et al., 2020; Granata, 2019; Granata et al., 2020), (4) satellite remote sensing (RS)-based models (Martens et al., 2017; Mu

* Corresponding author at: State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, College of Hydrology and Water Resources, Hohai University, Nanjing 210098, China.
E-mail address: qyduan@hhu.edu.cn (Q. Duan).

https://doi.org/10.1016/j.jhydrol.2021.126021
Received 21 November 2020; Received in revised form 14 January 2021; Accepted 18 January 2021
Available online 29 January 2021

et al., 2007), (5) land surface model (LSM) or hydrologic model simulation (Rodell et al., 2004; Srivastava et al., 2020; Sun et al., 2017; Xia et al., 2012; Zhang et al., 2014). There is no consensus on which method is best, as each category has its own advantages and disadvantages. Actually, there are considerable overlaps among the categories. For example, machine learning methods could serve as one kind of upscaling methods (Jung et al., 2011), while the machine learning methods in the third category are usually served as data-driven models to estimate ET with input forcing data. Satellite RS-based ET estimates rely on empirical or physical equations, which use the related satellite observations to calculate ET. The Penman–Monteith (PM) equation (Monteith, 1965) is a widely used satellite RS-based physical model (Mu et al., 2007, 2011), and it is also embedded in many LSMs and hydrologic models to compute ET (Kumar et al., 2017). Moreover, these empirical or physical equations can be also used alone with climatic forcing data (Deb et al., 2019; Zhang et al., 2019).

Among the above-mentioned methods, LSMs simulated ET has been receiving increasing attention due to its unique merits (Sun et al., 2017). LSMs can produce long-term gridded ET datasets with relatively higher spatial–temporal resolution at regional to global scales. In addition to ET, LSMs can be used to provide a consistent set of products of all fluxes and state variables under physical constraints (Zhang et al., 2014). LSMs also have the flexibility to use various in-situ observations and satellite RS data. Some intercomparison studies have demonstrated that the offline land surface modeling when forced with high-quality observations, can provide better ET estimates with smaller uncertainties compared to RS-based ET products and other kinds of ET products (Long et al., 2014; Wang et al., 2015; Xu et al., 2019). It should be noted that significant progress has been made in the development of regional and global land data assimilation systems (LDASs) based on LSMs. The LDASs aim at producing quality-controlled, long-term, spatially and temporally consistent fields of land surface states and fluxes by ingesting the best available observations (Rodell et al., 2004).

Despite the continuing efforts to improve the accuracy of LSM simulations, the errors of LSM-based ET estimates still cannot be neglected. Therefore, comprehensive evaluations are essential for better understanding the uncertainties, expanding their applications and motivating the model development. There have been numerous studies (Bai et al., 2018; Khan et al., 2018; Long et al., 2014; Mueller et al., 2011; Peters-Lidard et al., 2011; Xia et al., 2015; Xu et al., 2019; Zhang et al., 2020) evaluating LSM-based ET estimates, especially the ET products from the North American LDAS (NLDAS; Mitchell et al., 2004; Xia et al., 2012) and the Global LDAS (GLDAS; Rodell et al., 2004) projects. These evaluations are typically conducted at point scale using ET from EC measurements and at basin scale using the water balance method derived ET. However, EC systems are susceptible to the energy imbalance problem, which could cause errors in the measurements (Wilson et al., 2002; Xu et al., 2017). In addition, the EC sites are usually scarce, and the observed data is generally only available for a short time period with quite a few missing records. Due to the distinct spatial heterogeneity of ET, the applicability of ET site observations is very limited for large-scale evaluations. The water balance method is regarded as a reliable method to provide accurate long-term ET estimates at basin scale, however, it cannot represent the spatial variability of ET within the basin and its applicability is limited for small basins and relatively short time scales (e.g., daily). In addition, uncertainty in water balance method derived ET can even be higher than LSM-based ET (Long et al., 2014; Wang et al., 2015).

In addition to the limitations of the conventional ET evaluation methods mentioned above, the distributed, process-based LSMs hardly increase our process understanding if only evaluated at point scale or basin scale. Moreover, the evaluation results may be misleading for understanding the spatial predictability of LSM estimated ET. In fact, the scientific community has been advocating the spatial pattern evaluation of distributed models using spatial observation data for almost two decades (Beven and Feyen, 2002; Grayson et al., 2002; Refsgaard, 2001; Wealands et al., 2005). Recently, some innovative bias-insensitive spatial performance metrics have been applied to facilitate meaningful comparisons of spatial patterns of ET (Mendiguren et al., 2017; Koch et al., 2017) and other land surface variables (Fang et al., 2015; Koch et al., 2015, 2016), which could help to better assess and understand the spatial variability of land surface processes.

China stretches across a vast area covering a variety of climate regions and ecosystems and is facing many water issues. The LSMs in GLDAS can produce long-term, global ET products with relatively high resolution (up to 0.25° and 3 h) in near-real time. Along with ET, the global land surface fields provided by GLDAS could serve as important reference to support weather and climate prediction, water resources management, and water cycle studies in China. Recently, the new and reprocessed GLDAS Version 2 (GLDAS-2) data products have been released. Even though the old version GLDAS ET products have been evaluated over China in some previous studies (Bai et al., 2018; Ma et al., 2019; Wang et al., 2016), the conventional evaluation methods conducted at point scale or basin scale cannot reflect the large spatial heterogeneity of ET over China, thus reducing the generalizability of the conclusions. In another word, the traditional evaluation paradigm can no longer keep up with the continues progress of using advanced LSMs to develop large-scale ET estimates. To our best knowledge, there has been no study on spatial evaluation of LSM modeled ET over China yet. Therefore, the core novelty of this study is that, for the first time, a comprehensive spatial evaluation of the newly released GLDAS-2 ET products is performed over China. The spatial evaluation is based on three innovative bias-insensitive spatial performance metrics including the empirical orthogonal function (EOF) analysis, the connectivity analysis (Renard and Allard, 2013) and the fractions skill score (FSS; Roberts and Lean, 2008). In addition, this study also uses the EC measurements at 8 flux sites to evaluate the ET products. The main aims of this study are 1) to assess the performance of the LSMs in the upgraded GLDAS for estimating ET and provide reference for further development of model parameterization schemes and calibration methods; 2) to investigate the unique advantages of the spatial performance metrics which can enrich the ET evaluation approaches of the land surface and hydrological modeling communities.

The following section describes the GLDAS-2 ET products, the EC-based ET observations and ET data used as reference for spatial evaluation (section 2). Section 3 gives a detailed introduction of the evaluation methods. Section 4 presents comprehensive evaluations of the ET products from three LSMs in GLDAS. Section 5 presents detailed discussion of potential causes of the model deficiencies diagnosed by the spatial evaluations, followed by conclusions in Section 6.

2. Data

2.1. GLDAS ET products

GLDAS was developed jointly by the National Aeronautics and Space Administration (NASA) Goddard Space Flight Center (GSFC) and the National Oceanic and Atmospheric Administration (NOAA) National Centers for Environmental Prediction (NCEP). The goal of GLDAS is to generate optimal fields of land surface states and fluxes, by ingesting satellite- and ground-based observational data products, using advanced land surface modeling and data assimilation techniques (Rodell et al., 2004). GLDAS drives multiple, offline LSMs, integrates a huge quantity of observation-based data, and executes globally at high resolutions (2.5° to 1 km), enabled by the Land Information System (LIS; Kumar et al., 2006). Currently, GLDAS includes four LSMs: Noah (Chen et al., 1996; Koren et al., 1999), the Catchment LSM (CLSM; Koster et al., 2000), the Community Land Model (CLM; Dai et al., 2003), and the Variable Infiltration Capacity (VIC; Liang et al., 1994) model.

GLDAS-2 has three components: GLDAS-2.0, GLDAS-2.1, and GLDAS-2.2. GLDAS-2.0 is forced entirely with the Princeton meteorological forcing input data and provides a temporally consistent series

from 1948 through 2014. GLDAS-2.1 is forced with a combination of model and observation data from 2000 to present. GLDAS-2.2 products use data assimilation, whereas the GLDAS-2.0 and GLDAS-2.1 products have no data assimilation. The latest GLDAS-2 has made major adjustments since November 2019. First, with the upgraded LSMs and updated forcing data sets, the GLDAS-2.1 main production stream serves as a replacement for the old GLDAS version 1 (GLDAS-1) products, which were decommissioned in June 2020. Second, the GLDAS-2.0 Noah products were reprocessed with updated Princeton Forcing V2.2 Data and an upgraded version of Noah model (V3.6) in November and December 2019. In September 2020, GLDAS-2.0 VIC and CLSM products were publicly released. Third, GLDAS-2.2, which explores the data assimilation capabilities in the LIS, is new to the data archive. The GLDAS-2.2 products from CLSM-F2.5 with the Gravity Recovery and Climate Experiment (GRACE) data assimilation were released in February 2020.

The main objective of GLDAS-2.1 is to provide up-to-date global land surface model outputs, while preserving consistency of the long-term climatology (i.e., GLDAS-2.0) to the extent possible. Therefore, the GLDAS-2.1 ET products from Noah-3.6, CLSM-F2.5 and VIC-4.1.2 are evaluated in this study. All the GLDAS-2.1 products extend from 2000 to present with 3-hourly temporal resolution. The ET product from Noah has a spatial resolution of $0.25^\circ \times 0.25^\circ$, while the other two have spatial resolutions of $1^\circ \times 1^\circ$.

2.2. Eddy covariance measurements of ET

Nowadays, EC measurements of water vapor exchange and carbon dioxide are being made routinely on each continent. Since EC measurements can provide relatively accurate estimation of ET at a given site, the EC-based ET measurements of 8 flux measurement sites from ChinaFLUX are used for evaluation. ChinaFLUX is an observation and research network that applies EC and chamber methods to measure the exchanges of carbon dioxide, water vapor and energy between terrestrial ecosystem and atmosphere in China (Yu et al., 2006). ChinaFLUX has become an important part of the global network of flux measurement sites called FLUXNET (Baldocchi et al., 2001) and the main part of the regional network AsiaFlux.

Currently, the ChinaFLUX network includes 79 sites consisting of 18 cropland sites, 19 grassland sites, 23 forest sites, 15 wetland sites, 2 desert sites, 1 urban site and 1 waterbody site, which encompass a large range of latitudes, altitudes, climates and ecosystem types. Among the 79 sites, the flux measurements of the 8 sites used in this study are publicly available from http://www.chinaflux.org/. The spatial distribution of the 8 stations is shown in Fig. 1, while Table 1 lists their detailed information. The provided daily latent heat flux data at the 8 sites are used to derive the actual ET in units of water depth following Mu et al. (2011). In this study, we perform the grid-site comparison, meaning that the ET values of the GLDAS grid and the flux site that locates in that grid are directly compared at the same time scale (daily and monthly).

2.3. ET reference data across China

To perform a qualified spatial pattern evaluation of modeled ET, a reliable ET reference dataset is a prerequisite. Recently, Ma et al. (2019) developed a long-term (1982–2017) monthly terrestrial ET product with spatial resolution of 0.1° across China. This dataset is derived from a recently proposed nonlinear complementary relationship (CR) formulation (Szilagyi et al., 2017). The CR method, as first introduced by Bouchet (1963), emphasizes the feedback mechanism between actual ET and potential ET under the same environmental conditions. The CR method has been regarded as an attractive tool for estimating actual ET at large scale due to its minimal data requirement of only meteorological input. Interested readers can refer to Ma et al. (2019) for detailed information of this dataset, which is available from the National Tibetan Plateau Data Center (DOI: https://doi.org//10.11888/AtmosPhys.tpe.249493.file).

Ma et al. (2019) conducted independent evaluation based on EC measurements and water balance method estimated ET, the results indicated that the CR-based ET product was very reliable. Further evaluations suggested that it showed improved accuracy over seven other mainstream ET products. Therefore, this CR-based ET product can serve as a suitable reference for our spatial evaluation. The GLDAS-2.1 ET products are resampled from its original resolutions to 0.1° through bilinear interpolation to provide consistency with the CR-based ET product.

3. Methodology

3.1. Common statistic metrics

Three statistic metrics are used to conduct the grid-site evaluations of GLDAS ET products against EC measured ET in this study. The metrics include relative error (RE), bias adjusted root-mean square error (aRMSE) and Kling-Gupta efficiency (KGE; Gupta et al., 2009). The

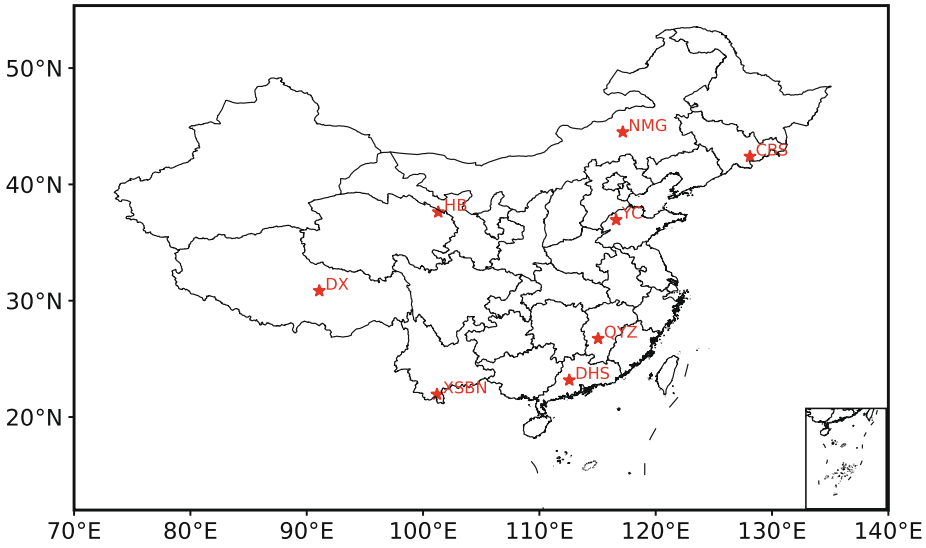

Fig. 1. The locations of the eight flux measurement sites.

Table 1
The information of the eight flux measurement sites used in this study.

Station	Location	Ecosystem type	Climate type	Elevation (m)	Data period
Changbaishan (CBS)	42.4°N, 128.1°E	Forest	Temperate continental monsoon climate	738	2003–2010
Qianyanzhou (QYZ)	26.74°N, 115.05°E	Forest	Subtropical monsoon climate	102	2003–2010
Dinghushan(DHS)	23.17°N, 112.57°E	Forest	Monsoon humid climate of torrid zone of south Asia	300	2003–2010
Xishuangbanna(XSBN)	21.95°N, 101.2°E	Forest	Monsoon humid climate of torrid zone of south Asia	750	2003–2010
Haibei (HB)	37.62°N, 101.31°E	Grassland	Highland continental climate	3250	2003–2010
Inner Mongolia (NMG)	44.5°N, 117.17°E	Grassland	Temperate semi-arid continental climate	1189	2004–2010
Dangxiong (DX)	30.85°N, 91.08°E	Grassland	Plateau monsoon climate	4333	2004–2010
Yucheng (YC)	36.95°N, 116.6°E	Cropland	Temperate semi-humid and monsoon climate	28	2003–2010

formulas are given by:

$$RE = \frac{\sum_{i=1}^{n}(S_i - O_i)}{\sum_{i=1}^{n} O_i} \times 100\% \quad (1)$$

$$aRMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[(S_i - \mu_s) - (O_i - \mu_o)]^2} \quad (2)$$

$$KGE = 1 - \sqrt{(r-1)^2 + (\beta - 1)^2 + (\gamma - 1)^2} \quad (3)$$

$$\beta = \frac{\mu_s}{\mu_o}$$

$$\gamma = \frac{\sigma_s}{\sigma_o}$$

where n is the total number of events; O_i and S_i are the ith pairs of EC-measured ET and GLDAS ET; μ_o and μ_s are their corresponding mean values. In the KGE equation, r represents the linear correlation coefficient, β and γ measure the bias and relative variability in the simulated and observed values. σ_s and σ_o are the standard deviation of simulated and observed variables, respectively.

Bias measures the average tendency of model simulations to over or under estimate the observations, and a value of 0 is perfect. The aRMSE is adopted to remove the systematic errors and show only the random errors. The KGE is designed to measure the Euclidian distance from the ideal point in the three-dimensional criteria space derived from the decomposition of Nash–Sutcliffe Efficiency (NSE). The optimal value of KGE is 1.

3.2. Spatial evaluation methods

The core content of this study is to perform true spatial evaluations of the GLDAS ET products based on three bias-insensitive spatial performance metrics, namely EOF analysis, FSS and connectivity analysis. These metrics or methods are not novel and were originally developed for other purposes. Not until recently were they introduced to facilitate meaningful spatial validation of land surface variables in the hydrological and land surface modeling community (Fang et al., 2015; Koch et al., 2015, 2016, 2017). To perform spatial evaluation, all the GLDAS-2.1 ET products are interpolated to 0.1°, the same resolution as the CR-based ET product. The detailed evaluation results are presented in Section 4, and some methodology limitations are discussed in Section 5.

(1) Empirical orthogonal function (EOF) analysis

The method of EOF, also known as principal component analysis (PCA) in statistics, is a decomposition of a dataset in terms of orthogonal basis functions. It is often used in atmosphere, climate, ocean, and hydrology science to study possible spatial modes of variability and how they change with time. The main feature of the EOF analysis is that it seeks structures that explain the maximum amount of variance in a two-dimensional dataset. Typically, one dimension in the dataset is the space dimension, and the other is the time dimension. The EOF decomposes a large spatiotemporal dataset into a set of mathematically orthogonal modes (structures), which are usually called EOFs and a set of time series (loadings) that are related one-to-one to the EOFs and quantify the amplitude of each EOF over the period of record. The first EOF explains the largest amount of the variance. More detailed description of the EOF method can be referred to Björnsson and Venegas (1997).

While the EOF analysis is typically employed for analyzing the spatiotemporal variability of a single filed, Koch et al. (2015) proposed the novel concept of performing a joint EOF analysis on a combined data matrix that contains both reference and modeled data. By doing so, not only do the obtained EOF maps represent the spatiotemporal variability of both datasets, but also the difference between the loadings at each time step can be used as an indicator of spatial similarity (Koch et al., 2016). To ensure a reliable pattern similarity score, the loading deviation must be weighted according to the variance contribution of the corresponding EOF. The EOF-based similarity score between a reference dataset and a modeled map at time t can be formulated as:

$$S_{EOF}^{t} = \sum_{i=1}^{n} w_i |(load_i^s - load_i^o)| \quad (4)$$

where w_i is the variance contribution of the ith EOF, n is the total number of orthogonal modes (EOFs), $load_i^s$ and $load_i^o$ are the corresponding loadings of simulated and observed maps, respectively. In this study, the monthly mean is removed from each ET map prior to the EOF analysis; thus, the method is based on the spatial anomalies which makes it bias insensitive.

(2) Fractions skill score

The FSS is a scale-selective method developed by Roberts and Lean (2008) to measure forecast skill of precipitation forecasts against spatial scale for a given threshold. It uses the concept of nearest neighbors as the means of selecting the scales of interest. This approach calculates the fractional coverage of binary events that have a value of 1 (have exceeded the threshold) in a given spatial window. Generally, percentile thresholds are used for conversion into a binary field to remove the impact of bias when focusing on the spatial accuracy. The main steps to obtain FSS are as follows: 1) convert the reference and the modeled spatial patterns into binary fields for a certain threshold; 2) for each grid in every binary field obtained in the last step, compute the fractions of grids with value of 1 within a given square window of length n; 3) calculate the mean square error (MSE) between the referenced and modeled fraction fields; 4) obtain the final FSS by normalizing the MSE from last step with the largest possible MSE that can be obtained from the modeled and referenced fractions. For a certain threshold, FSS at spatial scale of n is expressed as:

$$FSS_{(n)} = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N}(O_{(n)i} - S_{(n)i})^2}{\frac{1}{N}\left(\sum_{i=1}^{N} O_{(n)i}^2 + \sum_{i=1}^{N} S_{(n)i}^2\right)} \quad (5)$$

where $O_{(n)}$ and $S_{(n)}$ are the resultant fields of referenced fractions and modeled fractions, respectively. N is the total number of valid grids in the domain. The FSS ranges from 0 (complete mismatch) to 1 (perfect match). As the size of the square windows used to compute the fractions

becomes larger, the FSS will reaches an asymptote that depends on the ratio between the modeled and observed frequencies of the event. The FSS is bias insensitive due to the percentile thresholds used. Further information can be referred to Roberts and Lean (2008).

(3) Connectivity analysis

Connectivity analysis is usually used in hydrogeology to quantify aquifer heterogeneity, where the connectivity structure of the heterogeneity is understood as a property that strongly influences groundwater flow and solute transport (Renard and Allard, 2013). The studies of Western et al. 2001 and Grayson et al. (2002) were the early ones that applied the connectivity analysis to land surface variables. Connectivity can provide a reliable measure of the general structure and heterogeneity of spatial patterns, which can be used to evaluate the spatial performance (Koch et al., 2016; 2017). To apply this methodology to a field of continuous variable, Renard and Allard (2013) suggested the following steps: 1) decompose the field into a series of binary maps by introducing a set of increasing thresholds; 2) perform a cluster analysis on each of the binary map to identify connected clusters; 3) use the percolation theory to describe the transition from many disconnected clusters to a large connected cluster. Hovadik and Larue (2007) proposed to use the probability of connection as a suitable metric to quantify how percolated clusters are. The metric $\Gamma(t)$, which is computed for the threshold t, is formulated as:

$$\Gamma(t) = \frac{1}{n_t^2} \sum_{i=1}^{N(X_t)} n_i^2 \tag{6}$$

where n_t is the total number of grids in the binary map X_t, $N(X_t)$ represents the total number of distinct clusters, and n_i is the number of grids in the ith cluster. Like FSS, using percentiles makes this method bias insensitive, and facilitates separate studies of clusters of a binary map above or below a threshold. In this study, we use a series of thresholds which move along all percentiles (1%–100%) of the ET range. Therefore, the RMSE between the metrics of the reference field and modeled field for all thresholds can be computed to indicate the spatial similarity.

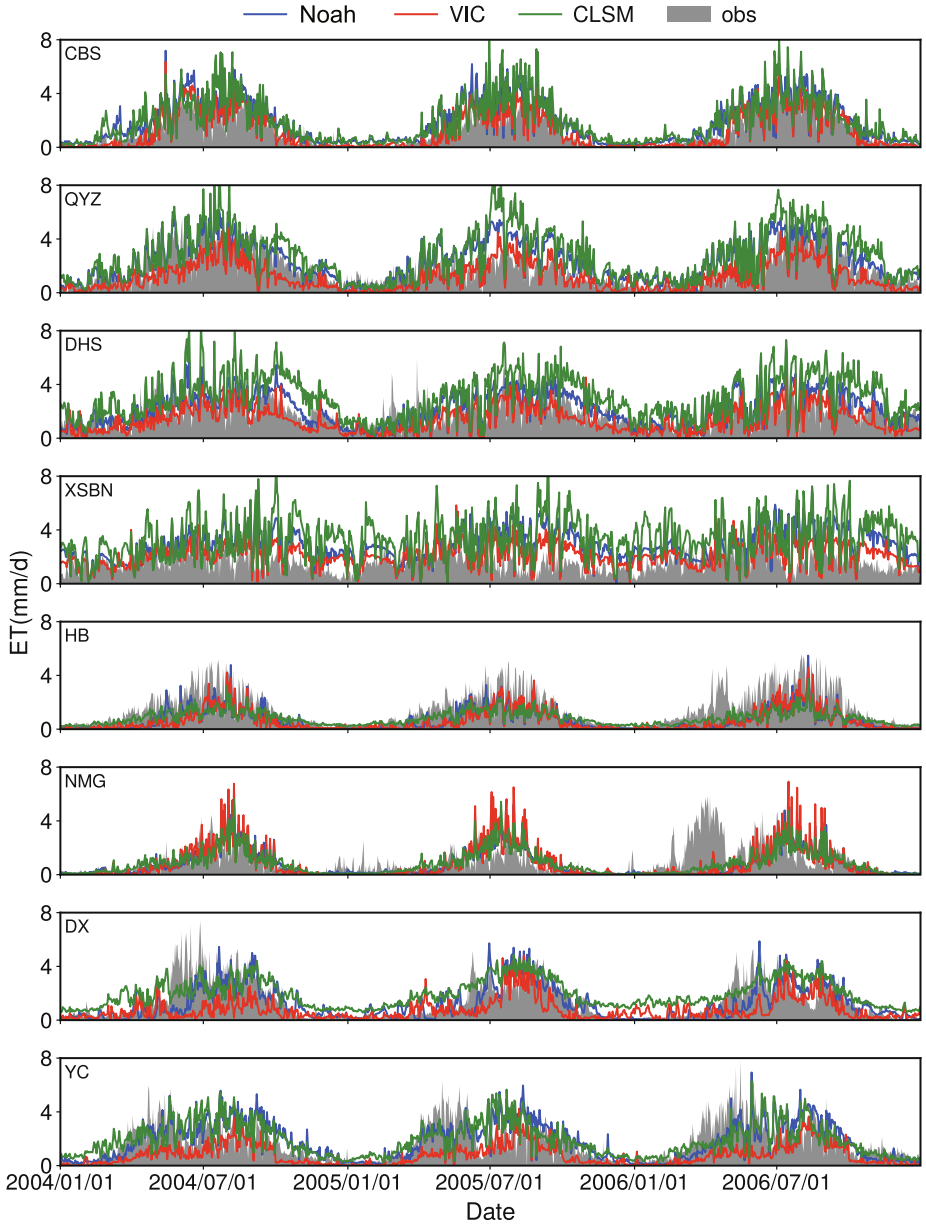

Fig. 2. Comparison of the daily ET estimates from the three LSMs with the ET measurements at the eight flux sites (only demonstrated over 2004–2006).

$$RMSE_{CON} = \sqrt{\frac{\sum_{t=1}^{100} (\Gamma(t)_o - \Gamma(t)_s)^2}{100}} \quad (7)$$

where $\Gamma(t)_o$ and $\Gamma(t)_s$ are the connectivity metrics of the reference and modeled field, respectively. Interested readers can refer to Renard and Allard (2013) and Koch et al. (2016) for further details.

4. Results

4.1. Grid-site evaluation

In this study, the three GLDAS-2.1 ET products are first validated against the EC measured ET of 8 flux sites on both daily and monthly time scales. In general, all three GLDAS products have the consistent trends with the observed ET series at all sites (Figs. 2 and 3). Moreover, the ET product from Noah performs well in capturing the magnitude at the QYZ and DX sites, and so does the VIC modeled ET at the CBS and DHS sites. However, a closer inspection shows that there are overall overestimations or underestimations for some GLDAS ET products at different sites, which are clearly demonstrated in Fig. 4. Specifically, Noah and CLSM modeled ET products show systematic positive biases at the four sites of the forest ecosystem, and CLSM is much worse, especially at the XSBN site. On the contrast, VIC shows systematic negative biases at two of the three grassland sites (HB and DX), the QYZ forest site as well as the YC cropland site.

To further quantitatively evaluate the performances of the three GLDAS ET products, we calculate three statistic metrics on both daily and monthly time scales (Tables 2 and 3). It can be found that no single LSM consistently outperforms the others at all sites. Overall, Noah and VIC have comparable performances while the former performs more stable, and they all outperform CLSM. Noah modeled ET product has the highest KGE values at more than half of the flux sites, while VIC modeled ET product has the superiority in producing lower aRMSE at more sites. Noah modeled ET product has the leading performance for at least one site of each ecosystem type. The highest KGE values for each ecosystem type are all higher than 0.6 and 0.7 (daily and monthly, respectively). VIC performs better at the sites of the forest ecosystem than those of

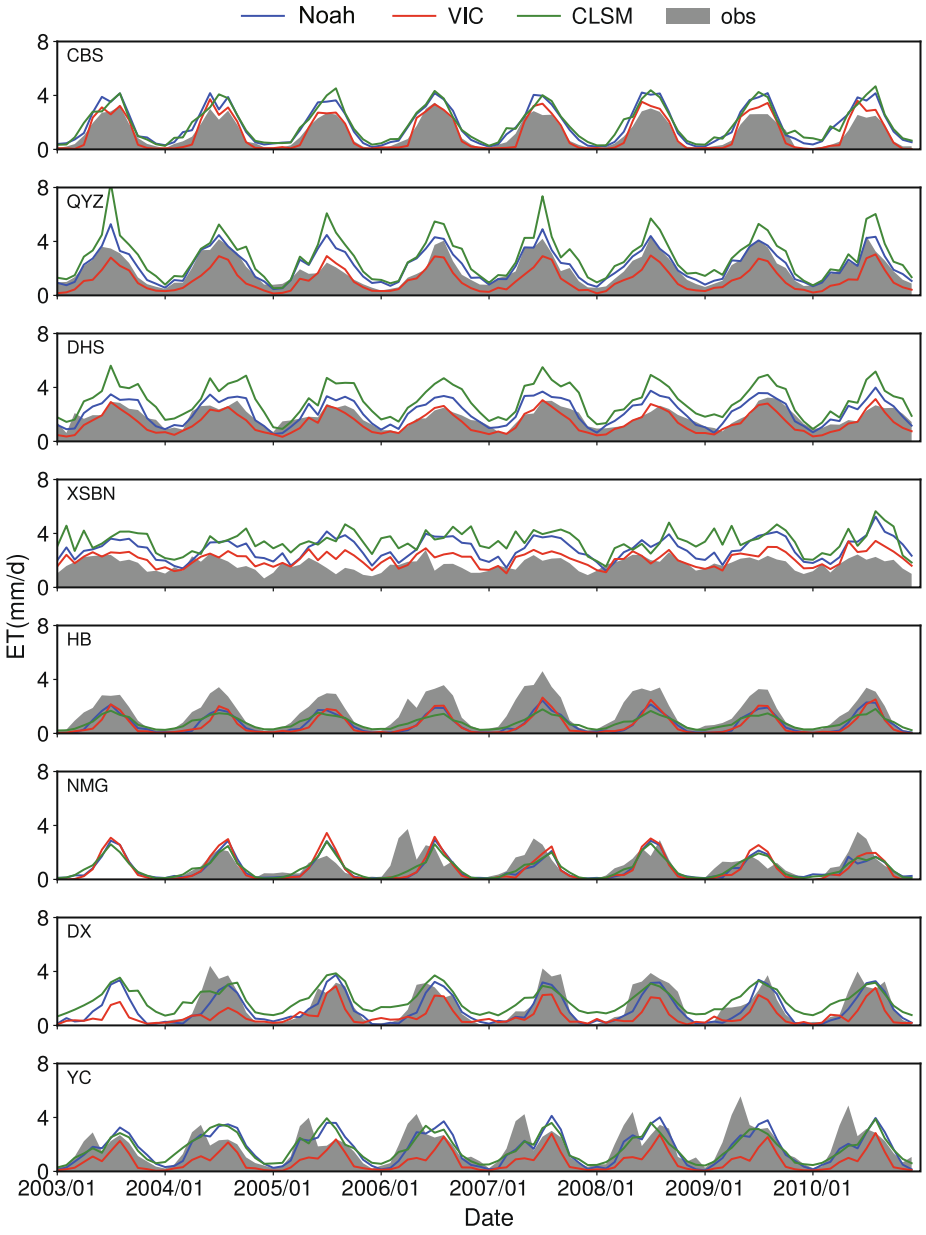

Fig. 3. Same as in Fig. 2, but for monthly ET over 2003–2010.

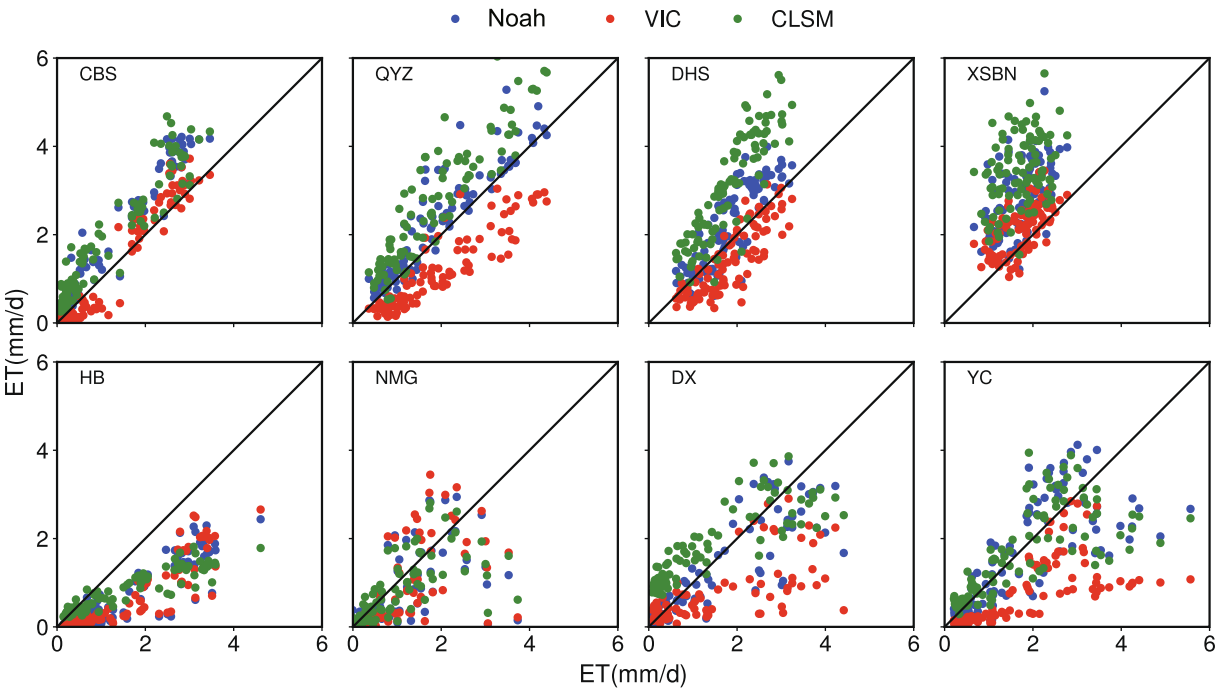

Fig. 4. Scatter plots showing the comparisons of the monthly ET estimates from the three LSMs with the ET measurements at the eight flux sites over 2003–2010.

Table 2
Statistical diagnostics of the daily ET estimates from the three LSMs at eight flux sites. The values shown in bold denote the best performance among the three LSMs for each site.

Station	RE (%)			aRMSE (mm/d)			KGE		
	Noah	VIC	CLSM	Noah	VIC	CLSM	Noah	VIC	CLSM
CBS	49.6	**−0.6**	55.5	0.79	**0.68**	0.93	0.36	**0.75**	0.31
QYZ	**17.3**	−38.3	47.8	**0.85**	0.94	1.20	**0.74**	0.46	0.35
DHS	**20.9**	−22.5	67.9	0.93	**0.87**	1.15	**0.57**	0.55	0.07
XSBN	61.4	**19.5**	93.7	1.18	**0.95**	1.56	0.04	**0.34**	−0.51
HB	−54.0	−58.4	**−47.8**	0.96	**0.95**	1.02	**0.29**	0.26	0.20
NMG	−24.3	−24.6	**−20.6**	1.04	1.17	**0.94**	0.40	0.39	**0.44**
DX	**−18.2**	−49.8	22.8	1.03	1.27	**1.01**	**0.64**	0.18	0.50
YC	−9.5	−57.7	**−7.2**	**1.22**	1.35	1.30	**0.63**	0.10	0.52

Table 3
Same as Table 2, but for monthly ET.

Station	RE (%)			aRMSE (mm/d)			KGE		
	Noah	VIC	CLSM	Noah	VIC	CLSM	Noah	VIC	CLSM
CBS	49.6	**−0.7**	55.5	0.42	**0.37**	0.46	0.42	**0.78**	0.39
QYZ	**17.3**	−38.3	47.7	**0.47**	0.49	0.80	**0.79**	0.55	0.34
DHS	**20.9**	−22.6	67.8	0.51	**0.41**	0.68	0.53	**0.70**	−0.02
XSBN	61.2	**19.4**	93.7	0.72	**0.46**	0.78	0.06	**0.46**	−0.24
HB	−54.1	−58.5	**−47.8**	**0.63**	**0.63**	0.78	**0.32**	0.30	0.21
NMG	−24.6	−25.0	**−20.9**	0.77	0.78	**0.68**	0.52	**0.56**	0.56
DX	**−18.3**	−49.9	23.0	**0.66**	0.99	0.74	**0.72**	0.24	0.56
YC	−9.6	−57.8	**−7.2**	**0.89**	0.96	0.92	**0.71**	0.21	0.62

other ecosystem types, while CLSM performs better at the grassland sites and the cropland site. It's noted that the metrics of aRMSE and KGE improve a lot when calculated on monthly time scale for nearly all LSMs and sites, indicating that the LSMs modeled ET products are more reliable on monthly time scale.

4.2. Spatial evaluation of ET patterns

Fig. 5 depicts the spatial distributions of multi-year (2000–2015) mean annual ET of the CR-based reference dataset and the three GLDAS ET products. In general, all of them present a consistent ET pattern with the increasing amount gradient from northwest to southeast. However, there are large differences in the mean annual values among the three GLDAS products and the reference, especially in the southeast China, the northwestern Tibetan Plateau and the Yangtze river basin. Among the three GLDAS LSMs, CLSM generally produces the highest ET values, followed by Noah and VIC. Fig. 6 further presents the seasonal mean ET patterns of the reference and the three LSMs during 2000–2015. In general, the patterns of spring, summer and autumn for the four ET products can reflect their corresponding mean annual ET patterns shown

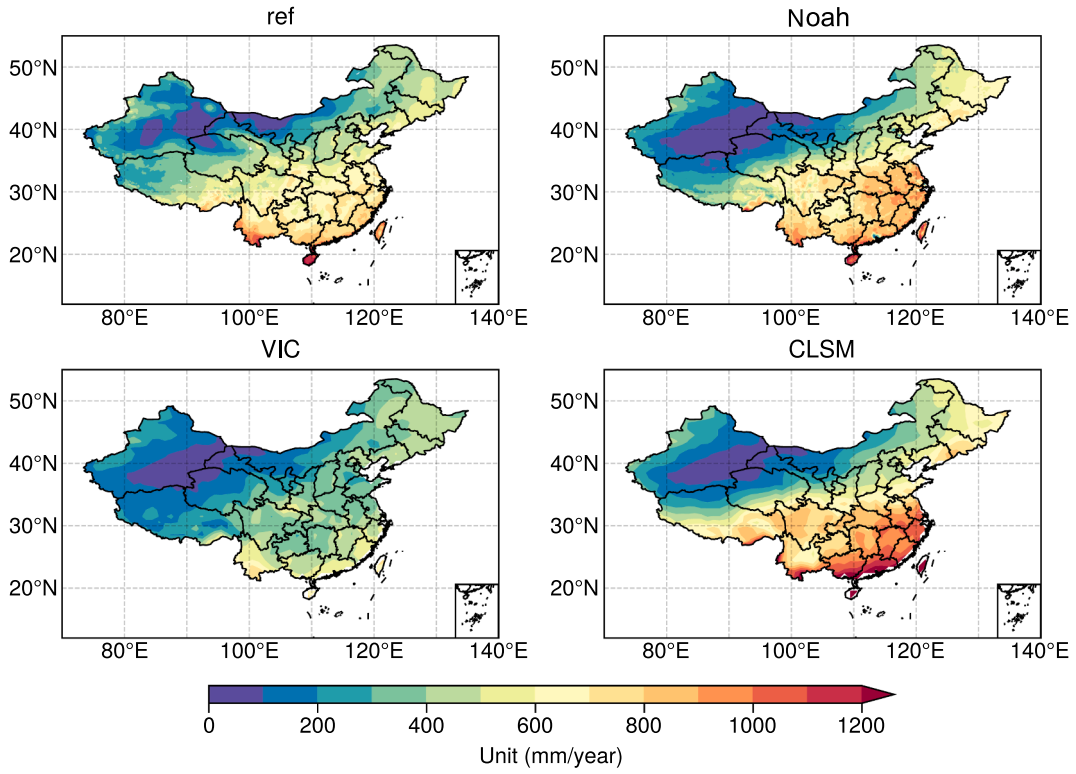

Fig. 5. The spatial distributions of multi-year (2000–2015) mean annual ET of the CR-based reference dataset and the three GLDAS ET products.

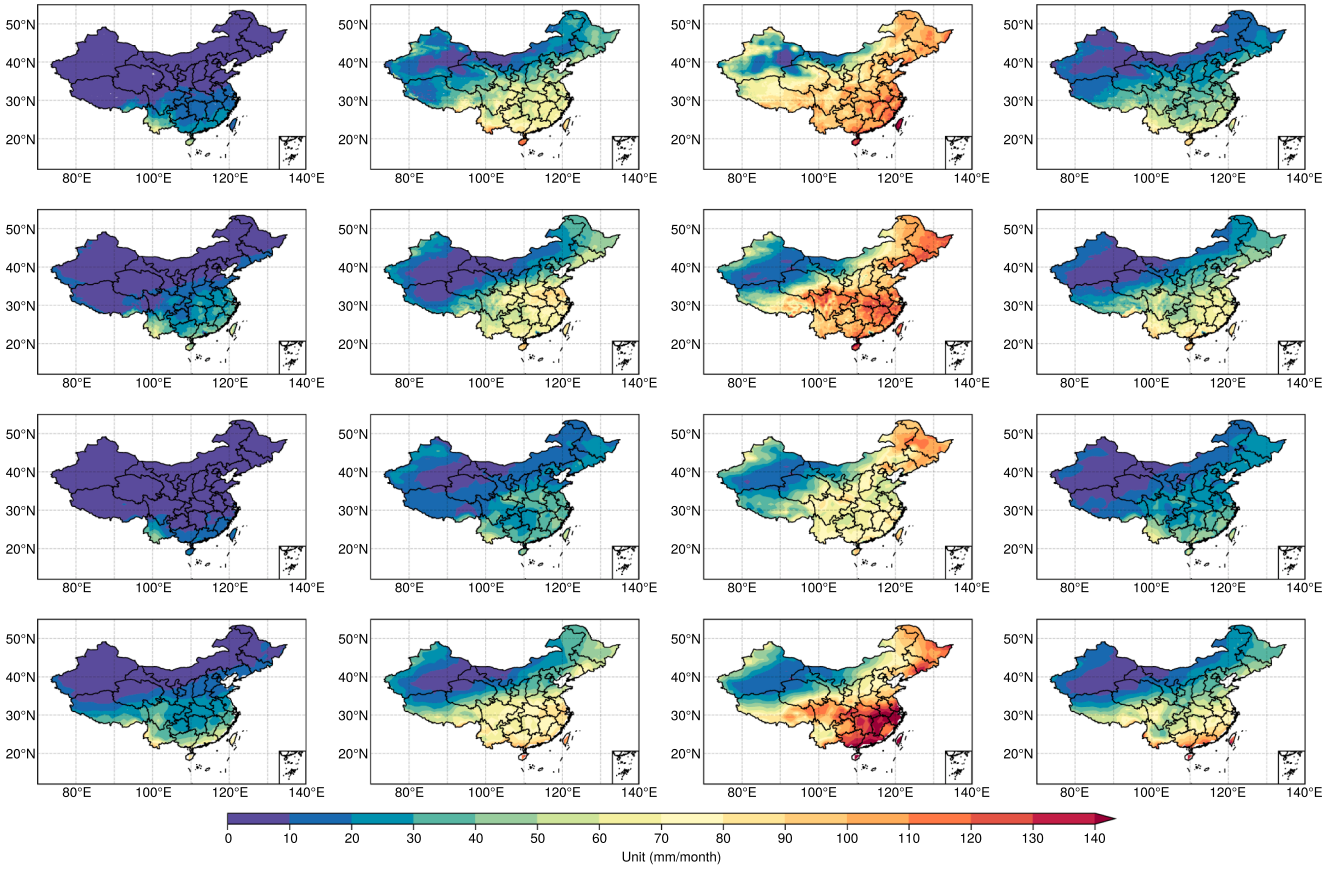

Fig. 6. The multi-year (2000–2015) seasonal mean ET patterns of the reference dataset, Noah, VIC and CLSM (from top to bottom) for winter, spring, summer and autumn (from left to right).

in Fig. 5. The winter ET patterns have the least apparent spatial heterogeneity due to the minimum precipitation and energy in this season. Among the three LSMs, VIC modeled ET patterns have the smallest spatial variability, especially for the winter season. In addition, CLSM produces the highest ET values across the seasons over vast parts of the northern China, whereas VIC modeled ET products have the lowest ET values.

From the above results we know that the three LSMs show different error characteristics. Nevertheless, different products may exhibit similar spatial patterns given their individual biases (Beck et al., 2017; Sun et al., 2016). Further evaluations are based on bias-insensitive metrics as the point-to-point evaluation may not be reliable.

a. EOF analysis

First, we conduct the joint EOF analyses for the CR-based reference dataset and each of the GLDAS ET product during the period of 2000–2015. Therefore, each EOF analysis is computed based on the decomposition of a concatenated spatiotemporal matrix of 384 monthly ET maps with spatial mean removed. Because of the mean removal, the EOF analysis is a bias-insensitive approach and not affected by the model bias. Each row in Fig. 7 shows the first two EOFs of the EOF analysis for Noah, VIC and CLSM, respectively. The first EOFs of Noah and CLSM can explain about 75% of the total variance, while the first EOF of VIC can explain about 65% of the variance. Moreover, the first two EOFs for all the GLDAS ET products can contribute more than 80% of the total variance. Generally, the values of the EOF maps do not represent the real amplitude of the original field. Instead, the first few EOF modes with large accumulated variance contribution are usually used to analyze the spatial characteristics of the original field, as well as covering the major information of the original field. The first EOF modes of the three GLDAS ET products are very similar and can depict the predominant pattern of the general increasing trend from northwest China to southeast China. In addition, the second EOF modes for the three LSMs also show very similar patterns, but the spatial variability is much more complex. It is notable that multiplying an EOF map with its corresponding loading can reflect the ET anomaly fields. Therefore, for all three GLDAS LSMs, the areas with positive values in EOF1 have negative deviations from the spatial mean, and vice versa, as the loadings corresponding to the first EOFs are negative all the time (Fig. 8). This result is consistent with Fig. 5, in which the lowest ET is observed in northwest China. Moreover, the first loadings also demonstrate an obvious characteristic of seasonal variation with peak values in summer and values close to zero in winter. This translates to large spatial variability in summer and small spatial variability in winter. The loadings for the second EOFs also show a strong seasonal signal, but the values switch from positive in winter to negative in summer, indicating that the pattern of the second EOF modes are inverted with seasons.

Fig. 8 also reveals that the loadings of the reference and three GLDAS ET products have consistent trends for both the first and the second EOF modes. The loadings of Noah and CLSM for the first EOF are lower than those of the reference, and the situation is opposite for VIC. This results in small spatial variability of VIC modeled ET compared to the reference dataset and the other two LSMs modeled ET products. The loadings of the three GLDAS ET products for the second EOF all show negative biases compared to those of the reference, and VIC has the largest biases among them. Fig. 9 further presents the EOF-based similarity scores for the three GLDAS ET products based on Eq. (4). Low values of the EOF-based metric are preferred, indicating the spatial performances of the modeled ET and the reference are very similar. All the time series of EOF-based scores show obvious seasonal signals with peaks in July and August and valleys in January. The small spatial variability in winter

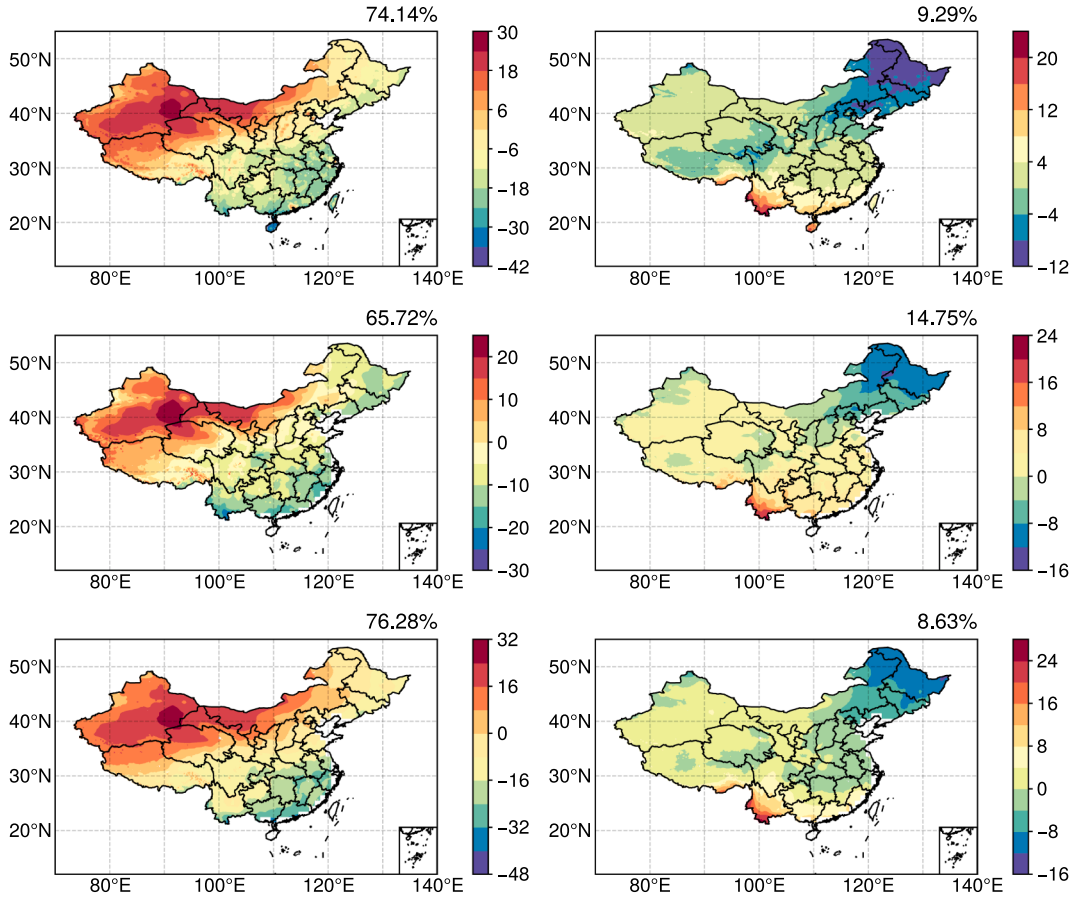

Fig. 7. The first (left column) and the second (right column) EOF modes of the joint EOF analyses for Noah, VIC and CLSM (from top to bottom).

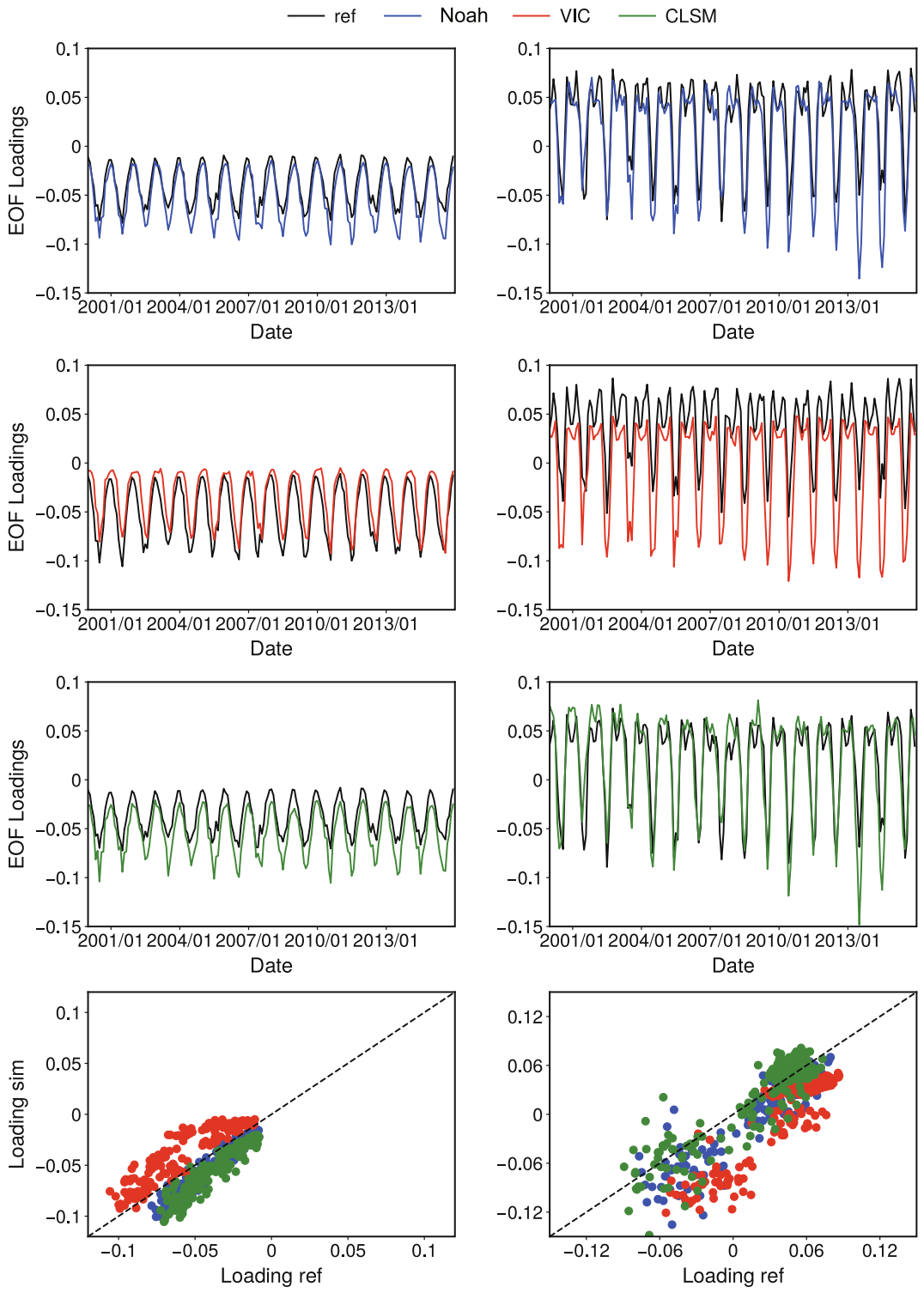

Fig. 8. Comparisons of the loadings of the reference dataset and those of the three LSMs modeled ET for the first (left column) and the second (right column) EOF modes.

makes it easier for the LSMs to reproduce the referenced spatial pattern. Fig. 9 reveals that the spatial performances of Noah and VIC are close in winter, which are also very similar to the reference. However, VIC provides a significantly worse spatial performance than Noah in summer. ET patterns modeled by CLSM are slightly worse than by Noah and VIC in winter, while they are better than those modeled by VIC in the hot months. Overall, Noah modeled ET has the best spatial performance in terms of the EOF-based metric among the three GLDAS ET products.

b. FSS

FSS is a scale-dependent verification method which has been widely used to evaluate precipitation forecasts. Recently, Koch et al. (2017) first applied this method to hydrological variables and stressed that it could clearly provide more information to a spatial pattern comparison. In our study, if a threshold percentile is below 50 the method will use the grids that fall below the value (low phase). On the contrary, it will focus on

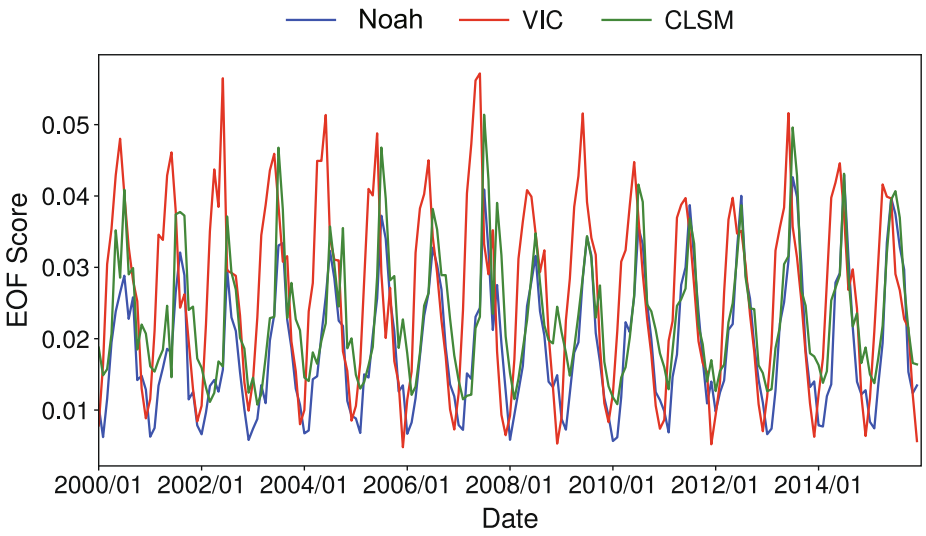

Fig. 9. Comparison of the EOF-based similarity scores for the LSMs modeled ET over 2000–2015.

grids exceeding the threshold (high phase) when it is above 50. Fig. 10 shows the FSS curves for the multi-year (2000–2015) mean ET patterns of July using different percentile thresholds. At all thresholds the FSS values increase gradually with the scale used to compute the fractions getting larger. This is the essential feature of the method because the numbers of grids with a value of 1 in the binary maps for the reference and model output are more likely to be close within a big square window. However, the increasing rate differs a lot among the thresholds. The differences in the spatial performance of the LSMs are clearly demonstrated by FSS. VIC and CLSM have very similar skill in capturing the patterns of low ET values. Noah performs much worse than VIC and CLSM for the 5% and 10% thresholds, but it achieves great improvement for the 20th percentile and has comparable spatial performance with the other two models. On the contrary, Noah has the best skill in reproducing the localized features in the regions of high ET values, and it greatly outperforms VIC for modeling the highest 5% and 10% ET

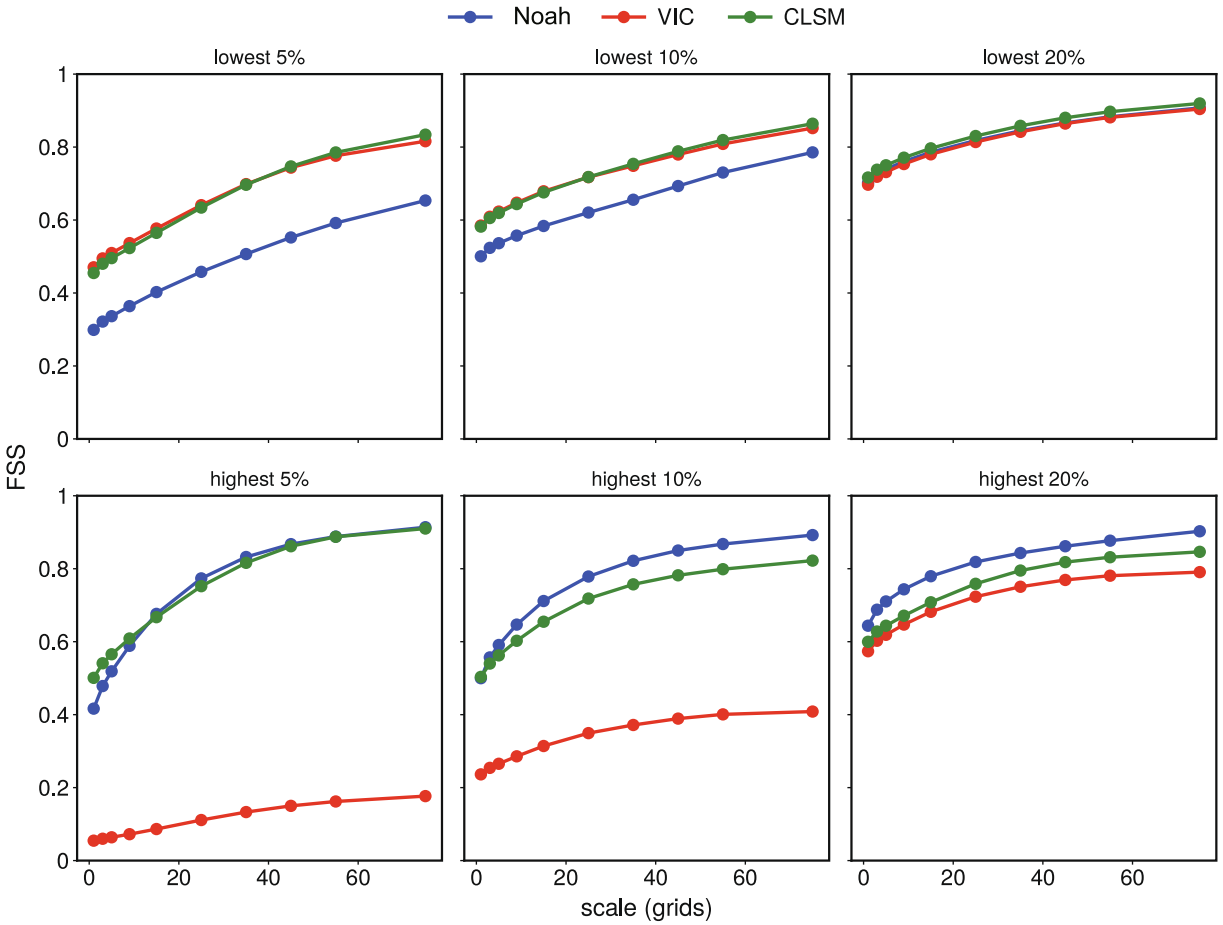

Fig. 10. Graphs of FSS against neighborhood length for the multi-year (2000–2015) mean ET patterns of July of the three LSMs using different percentile thresholds.

patterns. The reason why all the LSMs show a certain degree of improvement as the percentile moves toward the middle is that more localized ET features are more difficult to capture accurately.

Fig. 11 further depicts the FSS time series of the multi-year (2000–2015) average monthly ET patterns for the three GLDAS LSMs using different percentile thresholds. FSS is calculated at the predefined critical scale for each threshold following the suggestion of Koch et al. (2017). We use the critical scales of 65, 35 and 15 grids for calculating FSS at 5th, 20th and 40th percentiles (both top and bottom percentiles), respectively. The selection of critical scales is highly subjective, but we found that the results are basically the same with different selections of critical scales. For the low phase, FSSs of Noah and CLSM show distinct seasonality with lower values in cold months and higher values in warm months. VIC shows superiority in predicting the patterns of low ET in cold months. Noah produced FSS is the worst for the bottom 5th percentile, while it surpasses VIC in warm months for higher percentiles. For the high phase, VIC performs badly in capturing the localized patterns of the highest 5% ET in summer but performs well enough in wintertime. Noah and CLSM produced FSSs for the top 5% percentile feature large fluctuations but no obvious seasonality. Oppositely, Noah and CLSM provide good and consistent spatial performance and outperform VIC for the top 20th and 40th percentiles.

c. Connectivity analysis

The connectivity analysis is then applied to further assess and evaluate the spatial performance of the three LSMs. Fig. 12 gives an example of the cluster analysis of the reference and the three GLDAS ET products in July 2000. The first and second rows represent the connected clusters for the highest and the lowest 20% ET binary maps. For the top 20th percentile, the general cluster patterns of the three LSMs resemble that of the reference, but the differences in size and number of clusters are obvious. CLSM and Noah outperform VIC in terms of the spatial similarity of the cluster maps. The cluster of VIC in northeast China is much smaller than that of the reference, whereas VIC generates a much bigger cluster in northeast China. For the bottom 20th percentile, the three LSMs all reproduce the main cluster located in northwest China but miss the part in Inner Mongolia. In addition, they also fail to identify the small-sized clusters. Fig. 12 also shows the connectivity curves, which depict the probability of connection, $\Gamma(t)$, at all thresholds for the high (grid cells above the threshold) and low phase (grid cells below the threshold) of the referenced and modeled ET patterns in July 2000. As the percentile used as threshold increases for the high (from right to left in the x-axis) and low phase (from left to right in the x-axis), the connectivity $\Gamma(t)$ generally increases as well. Renard and Allard (2013) underlined that the percolation threshold, at which the connectivity increases abruptly, is a distinct characteristic of a spatial pattern. There is large disparity among the three LSMs in terms of their shapes of connectivity curves and percolation thresholds for the high phase. The connectivity curve of VIC shows the earliest percolation, indicating the overall large degree of homogeneity in it modeled ET pattern. However, VIC modeled ET pattern is more heterogeneous than the reference and other two LSMs at some thresholds lower than VIC's percolation threshold, for example, the highest 20% ET patterns shown in Fig. 12. Noah and CLSM simulated connectivity percolates later than the reference, and basically shows underestimation. This means that their modeled ET patterns are too heterogeneous relative to the reference. In contrast to the high phase, the three LSMs generate very similar connectivity curves for the low phase, which overestimate the connectivity and show very early percolations compared to the reference.

To investigate the features of the transition of different ET patterns from winter to summer, Fig. 13 illustrates the connectivity curves of the multi-year average ET patterns of February and August for the reference and the three LSMs. In general, the observed connectivity curves for the high phase show earlier percolations in February than in August, which is opposite to the low phase. The LSMs behave quite differently in terms of their connectivity and their difference is more distinct for the high phase. The RMSE between the connectivity curves as shown in Eq. (7) for both the high and low phase is used to quantify the pattern similarity with respect to the reference for each LSM modeled ET product. Fig. 14 shows the RMSE time series based on the connectivity analyses of the mean monthly ET patterns during 2000–2015 for the three LSMs. VIC's spatial performance for the high phase is worse than Noah and CLSM in warm months. Noah shows the largest spatial similarity to the reference in terms of the RMSE value in summer while CLSM has the best spatial performance in winter. For the low phase, the connectivity of the ET clusters shows similar RMSE values for the three LSMs except that VIC significantly outperforms Noah and CLSM in February and March.

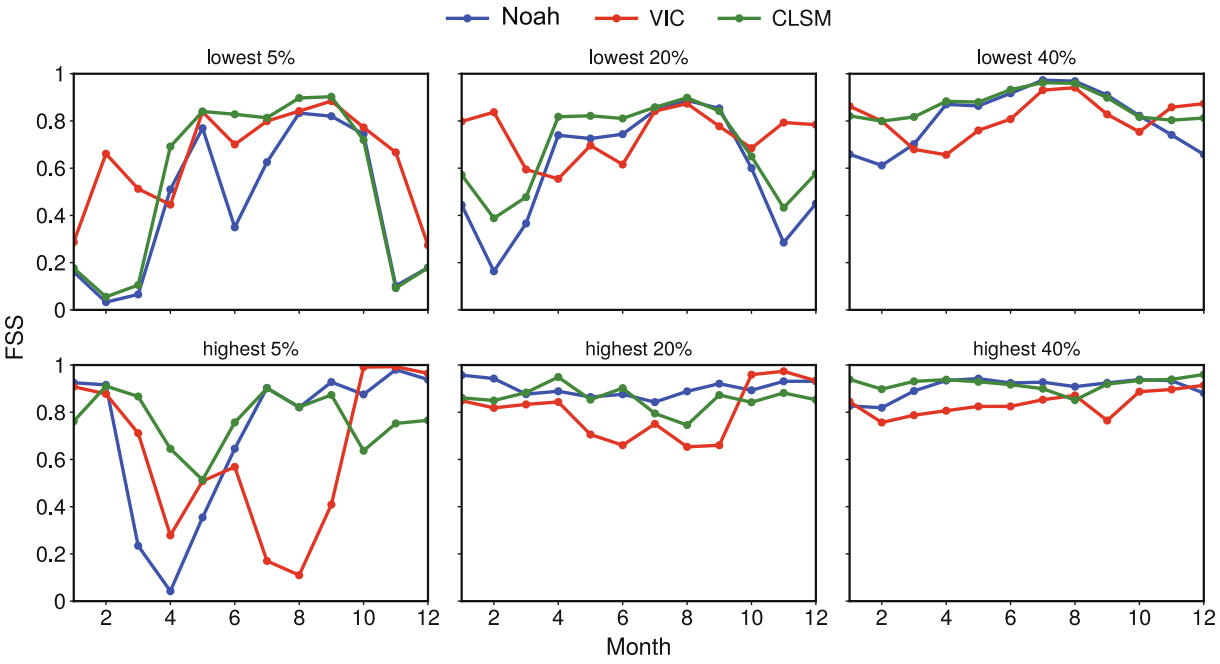

Fig. 11. Comparison of the FSS time series of the multi-year (2000–2015) average monthly ET patterns for the three LSMs using different percentile thresholds.

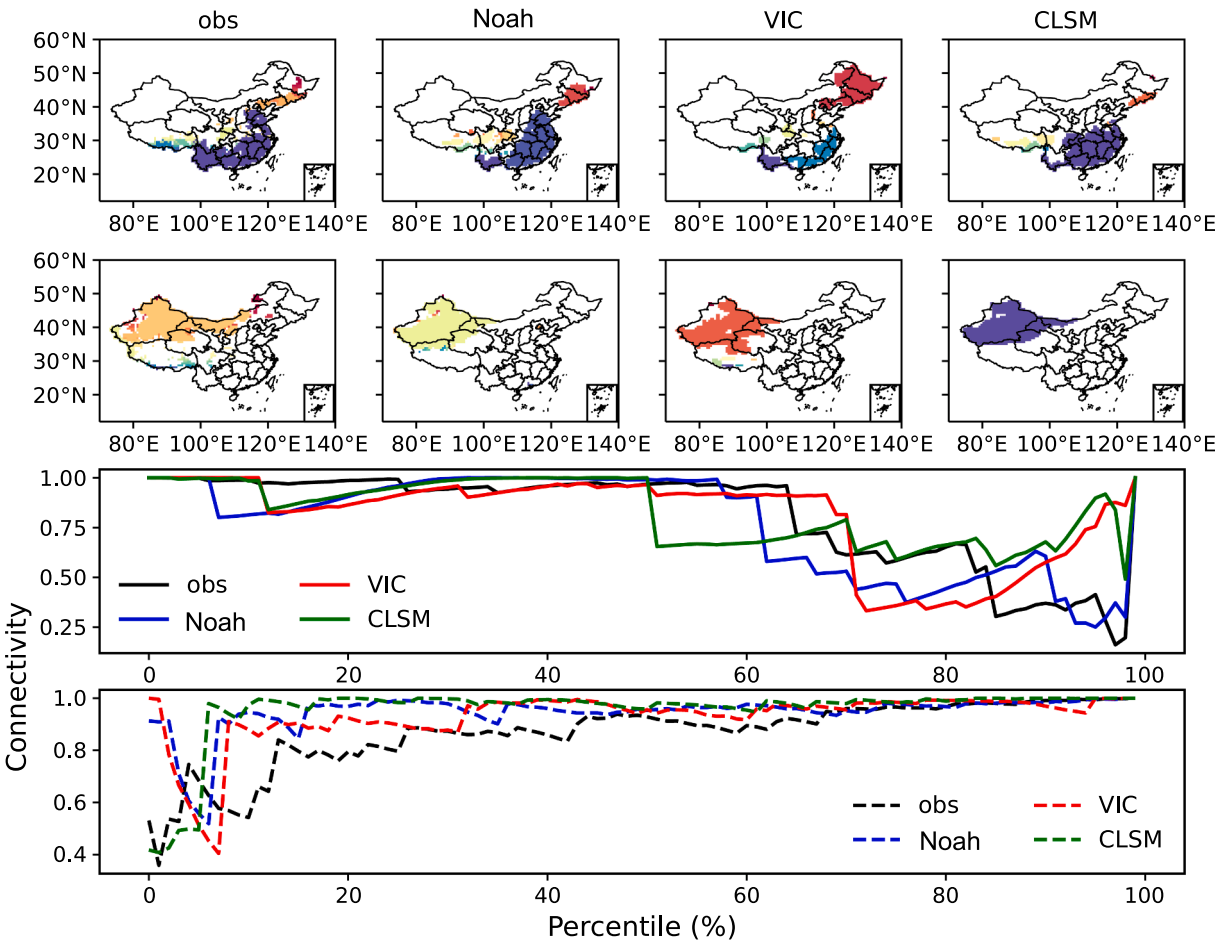

Fig. 12. The connected clusters of the highest 20% (first row) and the lowest 20% (second row) ET binary maps for the reference and the three GLDAS ET products in July 2000. The connectivity curves are shown at all percentile thresholds for the high phase (third row) and the low phase (fourth row).

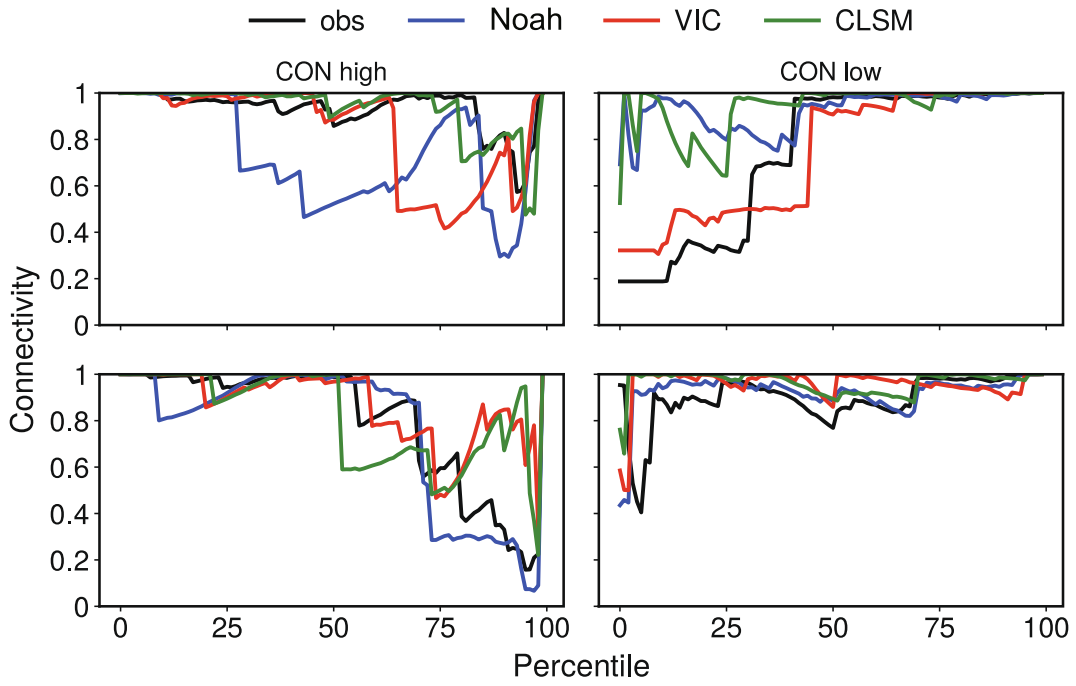

Fig. 13. The connectivity curves of the multi-year (2000–2015) average ET patterns of February (first row) and August (second row) for the high phase (left column) and the low phase (right column) corresponding to the reference and the three LSMs.

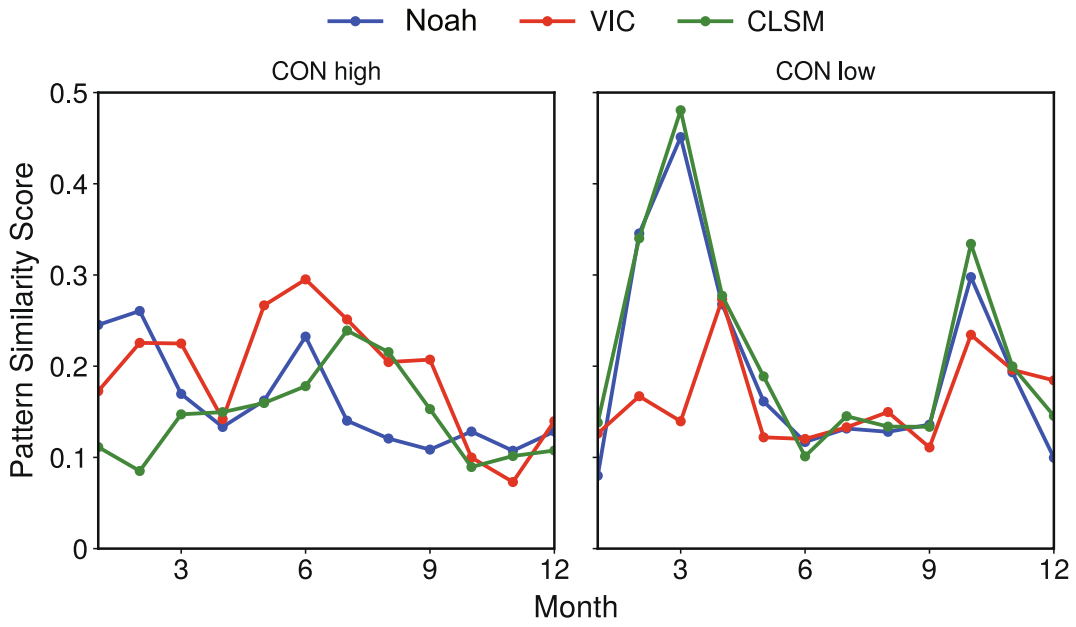

Fig. 14. The RMSE between the reference and model connectivity of the average monthly ET patterns during 2000–2015 for the high phase (left column) and the low phase (right column).

5. Discussion

5.1. Uncertainties in the evaluation methods

This study evaluated three reprocessed and recently released GLDAS-2.1 ET products based on the EC-measured ET values and the spatial evaluation methods. Although the EC flux measurements are considered as the most accurate method to provide ET estimates at a site, a direct grid-site comparison always suffers from the issues of spatial scale mismatch. For example, the highest spatial resolution of GLDAS ET products is 0.25°, which is far beyond the spatial representativeness of an EC flux site (Liu et al., 2011). In addition, EC systems are susceptible to the energy imbalance problem, and the number of EC flux sites is usually very limited. Therefore, the grid-site evaluation performed in this study only serves as an auxiliary part to help understand the errors of the new GLDAS ET products in general.

Spatially distributed and process-based modeling has been proliferated and regarded as a very important tool to provide predictions of the spatiotemporal variability of terrestrial energy, water and carbon cycle, which can help people to address a variety of environmental problems, such as climate change impacts, water resources management, drought monitoring. Therefore, the comprehensive spatial evaluation is of great importance to give model users and developers an insight into the spatial predictability of the models. The core content of this study is to conduct true spatial evaluations of the GLDAS ET products based on three bias-insensitive spatial performance metrics, namely EOF analysis, FSS and connectivity analysis. Although these metrics or methods are not novel, they were not used and focused on by the hydrological and land surface modeling community until recently.

The results show that CLSM performs worst based on the grid-site evaluation, whereas its spatial performance is overall better than VIC and even outperforms Noah sometimes. This indicates that the spatial performance of a model for representing the spatial variability of the natural system is not related to the model bias. In addition, no single method can offer enough information to quantify the spatial performance comprehensively. Different metrics may indicate opposite results as they provide different interpretations of the spatial similarity. For example, Noah performs the best in winter in terms of the EOF-based similarity score. However, the connectivity analysis of the high phase for Noah produces the worst RMSE performance, indicating the patterns are too heterogeneous compared to the reference. Therefore, a combination of metrics is suggested for a reliable spatial evaluation. Koch et al. (2017) underlined that the EOF analysis is the best option for a stand-alone metric. Connectivity analysis and FSS which compare spatial patterns at threshold percentiles can add unique information to a pattern evaluation besides the EOF analysis. However, one common drawback of the two threshold-based methods is that they may artificially increase spatial variability when the pattern is homogeneous. For example, the connectivity curves in Fig. 13 indicate that the patterns of low phase for the reference and VIC are very heterogeneous in February, which do not conform with the direct visual inspection of the ET patterns. After looking into the original data, we found that the connectivity analysis classified the grid cells with zero value and marginal values into different clusters, which led to the artificially increased spatial variability, especially for the reference and VIC.

5.2. Potential causes of the ET simulation errors

Even though the same meteorological forcing data are used to drive the LSMs in the GLDAS project, their ET simulations exhibit large differences regardless of the evaluation methods used. The differences are mainly attributed to the distinctions in model formulations of ET, model structures and model parametrizations. Generally, all LSMs use the PM approach for potential ET (PET) computations, then evaporation and transpiration are calculated by scaling PET (Kumar et al., 2018). The specific scaling method and model parameterizations differ in each LSM. A distinct feature of VIC is that it does not consider evaporation from soil underlying vegetations. In addition, the three LSMs also present differences in physics components such as soil hydrology, soil thermodynamics, and snowpack physics (Kumar et al., 2017). It should be noted that vegetation- and soil-related parameters also have large impacts on the modeled ET (Cuntz et al., 2016; Ma et al., 2019). Current GLDAS approach is to stay with model's default parameters which are indexed based on the soil texture classification and the vegetation classification. Noah uses the modified International Geosphere-Biosphere Programme (IGBP) 20-category vegetation classification based on the Moderate Resolution Imaging Spectroradiometer (MODIS) vegetation data, while VIC and CLSM use the University of Maryland (UMD) land cover classification. For the soil texture classification, Noah uses soil texture map is a hybrid of State Soil Geographic (STATSGO) over Continental United

States (CONUS) and Food and Agriculture Organization (FAO) soil map elsewhere, and the soil texture map for VIC and CLSM was derived from the soil fractions dataset from Reynolds et al. (2000), which were also based on the FAO Soil Map.

Although attributing the model deficiencies diagnosed by the spatial evaluations to specific causes is a difficult task, we still attempt to qualitatively account for the problems. We first looked into the model simulated transpiration, soil evaporation and canopy evaporation, which are the three components of the modeled ET. As shown in Fig. 15, VIC modeled soil evaporation is extremely small and shows no spatial variability compared to Noah and CLSM. Soil evaporation only accounts for a very small portion of the VIC modeled ET, which is consistent with the result of Kumar et al. (2018). As for the transpiration (Fig. 16) and canopy evaporation (not shown), VIC does not show significant differences in spatial pattern with Noah and CLSM. Because canopy evaporation is not the dominant source of ET, the spatial performance of VIC modeled ET is mainly determined by the modeled transpiration, which has smaller values and more homogenous patterns compared to Noah and CLSM modeled transpiration in southeast China, especially during summer. These factors caused the lack of spatial variability diagnosed by the EOF analysis, and the poor spatial performance reflected by FSS and the connectivity analysis in warm months, especially for the patterns of high phase. The diagnosed model deficiency of VIC is mainly attributed to the "big leaf" vegetation scheme in the VIC-4.1.2 model version, which was also indicated by Bohn and Vivoni (2016). It assumes there are no canopy gaps or exposed soil between plants, so soil evaporation only occurs in unvegetated areas. Since VIC-4.2 version, the "clumped'' vegetation scheme replaces the "big leaf" scheme. The former one divides each vegetation tile into vegetated and non-vegetated area fractions to account for soil evaporation and different wind and radiation attenuation in spaces between individual plants or gaps in the canopy (Bohn and Vivoni, 2016). The VIC-4.2 and later versions also support optional input of daily timeseries of LAI, albedo, and vegetated area fraction from forcing files instead of using the monthly climatology. Therefore, future GLDAS system could upgrade VIC to version 4.2 to improve the ET simulation. As for the good performance of VIC for the low phase in cold months pointed out by FSS and connectivity analysis, we know that it is mainly due to the artificially increased spatial variability based on the discussion in section 5.1. The similar phenomenon was also demonstrated by Koch et al. (2016) for spatial evaluation of VIC modeled LST over CONUS. Further investigation shows that the identified cluster patterns at extreme low percentiles in cold months are analogous to the snow water equivalent (SWE) pattern (not shown), which also has high similarity with the albedo pattern. In another word, VIC generates more snow cover areas and larger SWE in snow areas than the other two models, and the uncommon snow occurrence of VIC is strongly related to its extreme low ET values which are classified as a unique cluster. It's noted that the major update of VIC-4.1.2 was related to snowpack-related calculations and parameterizations (Xia et al., 2018), but the specific reason behind the phenomenon is worth further analysis.

The high connectivity of CLSM at extreme high percentiles indicates that the patterns of high ET values are too homogenous in summer, but its latest percolation stresses the overall heterogeneous pattern for the high phase. This phenomenon is mainly caused by the strong ET gradient across China with large transpiration areas located in southeast China. The significant larger transpiration simulated by CLSM in humid regions is possibly because CLSM develops vigorous upward diffusion of water to its root zone from its groundwater storage during the warm season (Mitchell et al., 2004), which is inherited from its predecessor-the Mosaic model (Koster and Suarez, 1992). In addition to transpiration, CLSM modeled soil evaporation is generally larger than that modeled by Noah and VIC. The overestimation of CLSM modeled ET in humid regions were also pointed out by some previous studies (Bai et al., 2018; Ma et al., 2019). CLSM employs a non-traditional approach where the subgrid heterogeneity of soil moisture in the root zone is statistically represented by separating the catchment into three distinct and dynamically varying subareas: (1) a saturated region where evaporation occurs without the consideration of water stress; (2) an unsaturated region where transpiration occurs with limited water stress and (3) a wilting region where transpiration is shut off. How this separation scheme impacts the calculation of soil evaporation deserves further study. As for the Noah model, it has overall better performance than VIC and CLSM in terms of both the grid-site evaluation and the spatial evaluations. However, Noah's spatial performance for high phase is inferior to CLSM and even slightly worse than VIC in late winter and early spring. The reason behind this remains ambiguous and needs to be investigated further in the future.

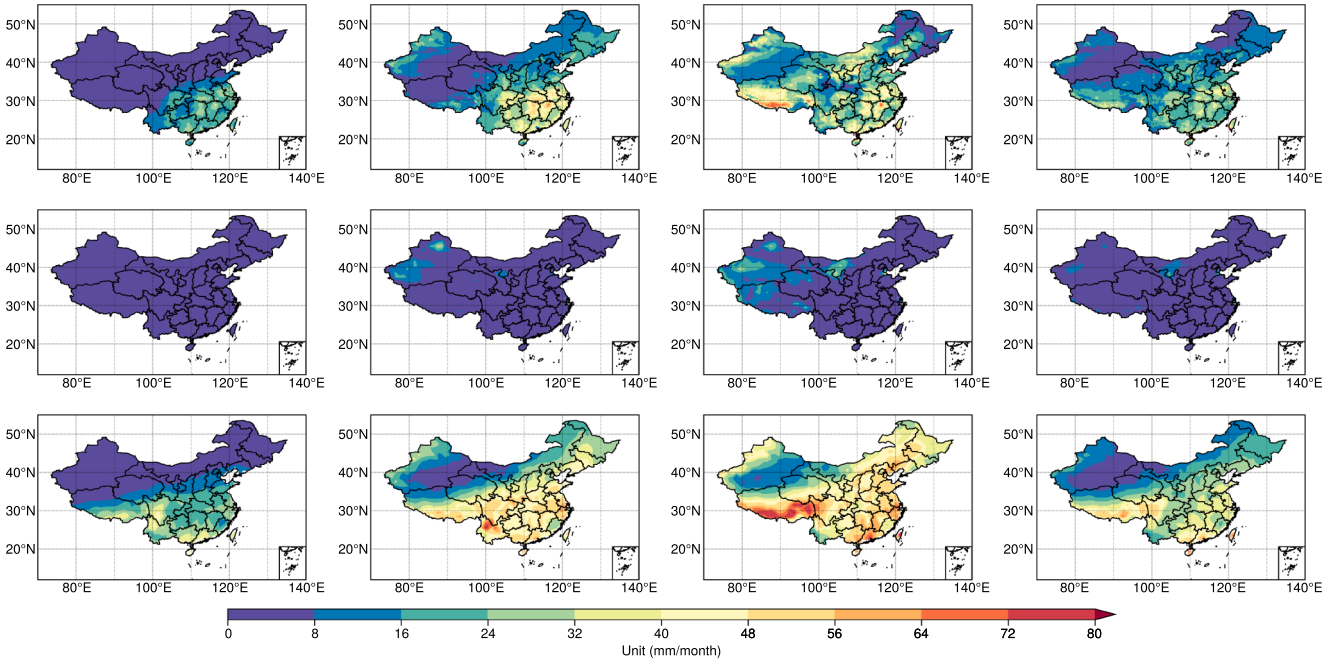

Fig. 15. Same as in Fig. 6, but for soil evaporation.

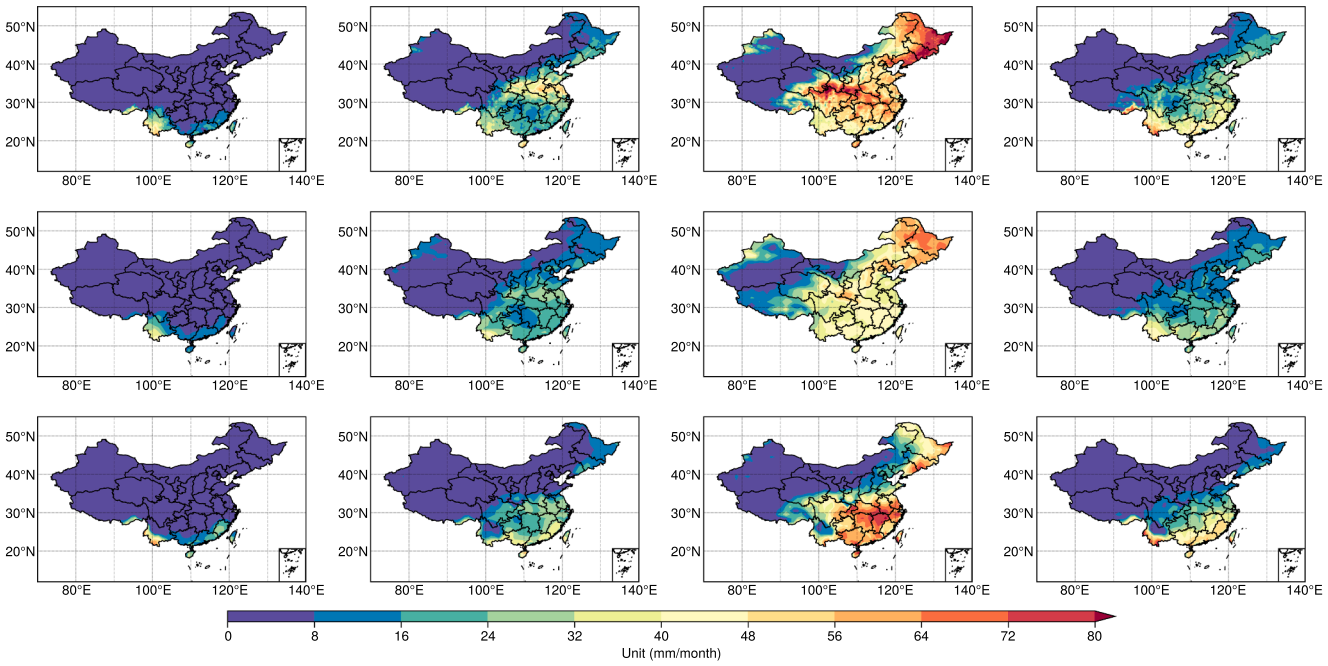

Fig. 16. Same as in Fig. 6, but for transpiration.

In addition to the respective deficiencies of the three LSMs, they all lack consideration of the effects of irrigation on ET. Moreover, they do not include a ground water module except CLSM. These two physical processes are very important for water cycle simulation and the lack of such processes may cause large simulation errors (Lawston et al., 2017; Xia et al., 2017, 2018). Another notable thing is that current GLDAS-2 models stay with their default parameter datasets as much as possible. However, in many cases the assigned default values based on land surface characteristics (e.g., soil and vegetation types) are inappropriate (Hou et al., 2012; Huang et al., 2013; Sun et al., 2020). Many studies have also highlighted the significant effects of model parameters on terrestrial processes modeling and the need of model calibration to improve them (Gong et al., 2016; Xia et al., 2018; Xu et al., 2019; Yang et al., 2016). Among all these potential causes of the differences and spatial deficiencies, it is difficult to tell which one plays a dominant role as they all contribute in some way to the uncertainty in modeling ET and may even compensate for each other. Therefore, future work is needed to systematically investigate the main drivers of spatial variability of the simulated ET and quantify the effects of different sources of uncertainties on land surface modeling.

Spatial evaluation of model is an important step during the process of model development. For example, Mendiguren et al. (2017) used the spatial evaluation method to reveal model structural insufficiencies and inconsistencies, which further helped to improve the model parameterizations. In addition, spatial pattern evaluation can also facilitate the development of alternative calibration strategies. Some recent studies have developed novel calibration methods of incorporating spatial pattern information of ET products to improve distributed hydrological modeling (Dembélé et al., 2020a, 2020b; Demirel et al., 2018). Therefore, the gained insights in this study will help us to develop long-term, high-resolution, and reliable ET products over China using LSMs with improved calibration methods and parameterization schemes, but it's beyond the scope of this study.

6. Conclusions

By ingesting the ground- and satellite-based observational datasets as well as taking advantages of the advanced techniques like data assimilation and model calibration, LSMs can produce long-term consistent ET products with high spatial–temporal resolution at regional to global scales, thus benefiting hydrometeorological research and applications. This study comprehensively evaluated the recently released GLDAS-2.1 ET products from three LSMs over China. We first performed a classical evaluation against EC-based ET observations at 8 sites. Then three spatial performance metrics which are relatively new to the land surface modeling communities were adopted to conduct the spatial evaluations of the GLDAS ET products. The primary conclusions are summarized as follows:

1. Evaluations against the EC measurements show that all three GLDAS products have the consistent trends with the observed ET series at both daily and monthly time scales. VIC tends to underestimate ET at most sites, while serious overestimations are observed in the simulated ET of CLSM. Noah and VIC have comparable performances in terms of different statistic metrics and outperform CLSM at both time scales.
2. The three GLDAS ET products could generally capture the spatial distribution of the ET reference dataset well. The main ET pattern with the increasing amount gradient from northwest to southeast China is consistent throughout the year. This dominant pattern is also captured by the joint EOF analyses of the referenced and simulated ET maps. The first EOFs of Noah and CLSM can explain about 75% of the total variance, while the first EOF of VIC can explain about 65% of the variance. In addition, the first two EOFs for all the GLDAS ET products contribute more than 80% of the total variance. The agreement between loadings of the reference and the simulated ET map is used to reflect the spatial performance quantitively. The EOF-based scores clearly indicate the spatial performances of the three LSMs are characterized by evident seasonal variations with the highest similarity in winter. Noah has the best performance in terms of the EOF-based score among the three LSMs.
3. Different spatial evaluation methods are not guaranteed to give consistent evaluation results because they focus on different aspects of the spatial performance. FSS and connectivity analysis, which are two spatial evaluation methods based on percentile thresholds, allow a separate analysis of the patterns of high phase and low phase. VIC generally performs the best in winter in terms of better performance of low phase and comparable performance of high phase with Noah

and CLSM. During warm months, VIC fails to reproduce the patterns of high phase, due to the low spatial variability indicated by the connectivity analysis. In contrast, VIC's performance for low phase is close to Noah and CLSM in warm months. Noah and CLSM show very similar spatial performance for representing the patterns of low phase. For high phase, Noah performs slightly better than CLSM in summer, mainly because that CLSM's patterns of high phase are too heterogeneous.

Overall, among the three GLDAS ET products, Noah modeled ET product shows certain superiority in matching the ET value and pattern. Thus, we recommend it as the first choice of the GLDAS ET products. This study shows that models with more accurate ET estimates at given sites may have worse spatial performance. The spatial evaluation methods thus show certain advantages over the traditional grid-site evaluation method of ET, which is valuable for large scale ET estimates, especially in the regions and countries where flux measurement sites are very scarce. Our study also indicates that the spatial evaluation methods can be used as effective tools to diagnose modeling deficiencies, thus helping future development and improvement of modeling techniques. More importantly, these methods are easily transferable to other distributed models and variables. Therefore, we recommend the spatial evaluation should be considered as a new paradigm in land surface and hydrological modeling communities.

Our future works will explore the main drivers of spatial variability of terrestrial ecological and hydrological process modeling. Moreover, further efforts are also needed to improve the model spatial predictability by developing high-quality model inputs, new parameterization schemes and calibration methods.

CRediT authorship contribution statement

Ruochen Sun: Conceptualization, Methodology, Software, Data curation, Formal analysis, Visualization, Investigation, Writing - original draft, Writing - review & editing. **Qingyun Duan:** Conceptualization, Supervision, Funding acquisition. **Jiahu Wang:** Data curation, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was jointly supported by the China Postdoctoral Science Foundation (2019M661714), the Fundamental Research Funds for the Central Universities (B200202031), the National Natural Science Foundation of China (51979004, 41830752), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA2006040104) and the National Key RD Program of China (2018YFC150806, 2019YFC1510700). The numerical calculations in this paper have been done on the computing facilities in the High Performance Computing Center (HPCC) of Nanjing University.

The GLDAS-2.1 ET products are available from the website https://disc.gsfc.nasa.gov/datasets?keywords = GLDAS. The EC measurements at 8 flux sites from ChinaFLUX can be downloaded from http://www.chinaflux.org/. The CR-based ET reference dataset is available from http://data.tpdc.ac.cn/en/.

References

Adnan, R.M., Chen, Z., Yuan, X., Kisi, O., El-Shafie, A., Kuriqi, A., Ikram, M., 2020. Reference evapotranspiration modeling using new heuristic methods. Entropy. 22, 547. https://doi.org/10.3390/e22050547.

Alizamir, M., Kisi, O., Muhammad Adnan, R., Kuriqi, A., 2020. Modelling reference evapotranspiration by combining neuro-fuzzy and evolutionary strategies. Acta Geophys. 68 (4), 1113–1126. https://doi.org/10.1007/s11600-020-00446-9.

Anderson, M.C., Allen, R.G., Morse, A., Kustas, W.P., 2012. Use of Landsat thermal imagery in monitoring evapotranspiration and managing water resources. Remote Sens. Environ. 122, 50–65. https://doi.org/10.1016/j.rse.2011.08.025.

Anderson, M.C., Hain, C., Otkin, J., Zhan, X., Mo, K., Svoboda, M., Wardlow, B., Pimstein, A., 2013. An intercomparison of drought indicators based on thermal remote sensing and NLDAS-2 simulations with U.S Drought Monitor Classifications. J. Hydrometeorol. 14, 1035–1056. https://doi.org/10.1175/JHM-D-12-0140.1.

Anderson, M.C., Kustas, W.P., Norman, J.M., Hain, C.R., Mecikalski, J.R., Schultz, L., González-Dugo, M.P., Cammalleri, C., D'Urso, G., Pimstein, A., Gao, F., 2011. Mapping daily evapotranspiration at field to continental scales using geostationary and polar orbiting satellite imagery. Hydrol. Earth Syst. Sci. 15, 223–239. https://doi.org/10.5194/hess-15-223-2011.

Bai, P., Liu, X., Liu, C., 2018. Improving hydrological simulations by incorporating GRACE data for model calibration. J. Hydrol. 557, 291–304. https://doi.org/10.1016/j.jhydrol.2017.12.025.

Baldocchi, D., Falge, E., Gu, L., Olson, R., Hollinger, D., Running, S., Anthoni, P., Bernhofer, C., Davis, K., Evans, R., Fuentes, J., Goldstein, A., Katul, G., Law, B., Lee, X., Malhi, Y., Meyers, T., Munger, W., Oechel, W., Paw, U.K.T., Pilegaard, K., Schmid, H.P., Valentini, R., Verma, S., Vesala, T., Wilson, K., Wofsy, S., 2001. FLUXNET: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. Bull. Am. Meteorol. Soc. 82, 2415–2434. https://doi.org/10.1175/1520-0477(2001)082<2415:FANTTS>2.3.CO;2.

Beck, H.E., Vergopolan, N., Pan, M., Levizzani, V., van Dijk, A.I.J.M., Weedon, G.P., Brocca, L., Pappenberger, F., Huffman, G.J., Wood, E.F., 2017. Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling. Hydrol. Earth Syst. Sci. 21 (12), 6201–6217. https://doi.org/10.5194/hess-21-6201-201710.5194/hess-21-6201-2017-supplement.

Beven, K., Feyen, J., 2002. The future of distributed modelling. Hydrol. Process. 16, 169–172. https://doi.org/10.1002/hyp.325.

Björnsson, H., Venegas, S.A., 1997. A manual for EOF and SVD analyses of climatic data. CCGCR Report. 97, 112–134.

Bohn, T.J., Vivoni, E.R., 2016. Process-based characterization of evapotranspiration sources over the North American monsoon region. Water Resour. Res. 52 (1), 358–384. https://doi.org/10.1002/wrcr.v52.110.1002/2015WR017934.

Bouchet, R.J., 1963. Evapotranspiration réelle evapotranspiration potentielle, signification climatique. IAHS Publ. 62, 134–142.

Chen, F., Mitchell, K., Schaake, J., Xue, Y., Pan, H.-L., Koren, V., Duan, Q.Y., Ek, M., Betts, A., 1996. Modeling of land surface evaporation by four schemes and comparison with FIFE observations. J. Geophys. Res. Atmos. 101 (D3), 7251–7268. https://doi.org/10.1029/95JD02165.

Cuntz, M., Mai, J., Samaniego, L., Clark, M., Wulfmeyer, V., Branch, O., Attinger, S., Thober, S., 2016. The impact of standard and hard-coded parameters on the hydrologic fluxes in the Noah-MP land surface model. J. Geophys. Res. Atmos. 121, 10676–10700. https://doi.org/10.1002/2016JD025097.

Dai, Y., Zeng, X., Dickinson, R.E., Baker, I., Bonan, G.B., Bosilovich, M.G., Denning, A.S., Dirmeyer, P.A., Houser, P.R., Niu, G., Oleson, K.W., Schlosser, C.A., Yang, Z.L., 2003. The common land model. Bull. Am. Meteorol. Soc. 84, 1013–1023. https://doi.org/10.1175/BAMS-84-8-1013.

Deb, P., Kiem, A.S., Willgoose, G., 2019. Mechanisms influencing non-stationarity in rainfall-runoff relationships in southeast Australia. J. Hydrol. 571, 749–764. https://doi.org/10.1016/j.jhydrol.2019.02.025.

Dembélé, M., Ceperley, N., Zwart, S.J., Salvadore, E., Mariethoz, G., Schaefli, B., 2020a. Potential of satellite and reanalysis evaporation datasets for hydrological modelling under various model calibration strategies. Adv. Water Resour. 143, 103667. https://doi.org/10.1016/j.advwatres.2020.103667.

Dembélé, M., Hrachowitz, M., Savenije, H.H.G., Mariéthoz, G., Schaefli, B., 2020b. Improving the predictive skill of a distributed hydrological model by calibration on spatial patterns with multiple satellite data sets. Water Resour. Res. 56, 1–26. https://doi.org/10.1029/2019WR026085.

Demirel, M.C., Mai, J., Mendiguren, G., Koch, J., Samaniego, L., Stisen, S., 2018. Combining satellite data and appropriate objective functions for improved spatial pattern performance of a distributed hydrologic model. Hydrol. Earth Syst. Sci. 22 (2), 1299–1315. https://doi.org/10.5194/hess-22-1299-2018.

Fang, Z., Bogena, H., Kollet, S., Koch, J., Vereecken, H., 2015. Spatio-temporal validation of long-term 3D hydrological simulations of a forested catchment using empirical orthogonal functions and wavelet coherence analysis. J. Hydrol. 529, 1754–1767. https://doi.org/10.1016/j.jhydrol.2015.08.011.

Fisher, J.B., 2014. Land-atmosphere interactions, evapotranspiration. In: Njoku, E.G. (Ed.), Encyclopedia of Remote Sensing. Encyclopedia of Earth Sciences Series. Springer, New York.

Fisher, J.B., Melton, F., Middleton, E., Hain, C., Anderson, M., Allen, R., McCabe, M.F., Hook, S., Baldocchi, D., Townsend, P.A., Kilic, A., Tu, K., Miralles, D.D., Perret, J., Lagouarde, J.-P., Waliser, D., Purdy, A.J., French, A., Schimel, D., Famiglietti, J.S., Stephens, G., Wood, E.F., 2017. The future of evapotranspiration: Global requirements for ecosystem functioning, carbon and climate feedbacks, agricultural management, and water resources. Water Resour. Res. 53 (4), 2618–2626. https://doi.org/10.1002/2016WR020175.

Gong, W., Duan, Q., Li, J., Wang, C., Di, Z., Ye, A., Miao, C., Dai, Y., 2016. Multiobjective adaptive surrogate modeling-based optimization for parameter estimation of large, complex geophysical models. Water Resour. Res. 52 (3), 1984–2008. https://doi.org/10.1002/2015WR018230.

Granata, F., 2019. Evapotranspiration evaluation models based on machine learning algorithms—A comparative study. Agric. Water. Manag. 217, 303–315. https://doi.org/10.1016/j.agwat.2019.03.015.
Granata, F., Gargano, R., de Marinis, G., 2020. Artificial intelligence based approaches to evaluate actual evapotranspiration in wetlands. Sci. Total Environ. 703, 135653. https://doi.org/10.1016/j.scitotenv.2019.135653.
Grayson, R.B., Blöschl, G., Western, A.W., McMahon, T.A., 2002. Advances in the use of observed spatial patterns of catchment hydrological response. Adv. Water Resour. 25, 1313–1334. https://doi.org/10.1016/S0309-1708(02)00060-X.
Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J. Hydrol. 377 (1-2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003.
Hou, Z., Huang, M., Leung, L.R., Lin, G., Ricciuto, D.M., 2012. Sensitivity of surface flux simulations to hydrologic parameters based on an uncertainty quantification framework applied to the Community Land Model. J. Geophys. Res. Atmos. 117 (D15) https://doi.org/10.1029/2012JD017521.
Hovadik, J.M., Larue, D.K., 2007. Static characterizations of reservoirs: Refining the concepts of connectivity and continuity. Petrol. Geosci. 13 (3), 195–211. https://doi.org/10.1144/1354-079305-697.
Huang, M., Hou, Z., Leung, L.R., Ke, Y., Liu, Y., Fang, Z., Sun, Y., 2013. Uncertainty analysis of runoff simulations and parameter identifiability in the Community Land Model: Evidence from MOPEX basins. J. Hydrometeorol. 14, 1754–1772. https://doi.org/10.1175/JHM-D-12-0138.1.
Jung, M., Reichstein, M., Margolis, H.A., Cescatti, A., Richardson, A.D., Arain, M.A., Arneth, A., Bernhofer, C., Bonal, D., Chen, J., Gianelle, D., Gobron, N., Kiely, G., Kutsch, W., Lasslop, G., Law, B.E., Lindroth, A., Merbold, L., Montagnani, L., Moors, E.J., Papale, D., Sottocornola, M., Vaccari, F., Williams, C., 2011. Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations. J. Geophys. Res. Biogeosciences. 116, 1–16. https://doi.org/10.1029/2010JG001566.
Katul, G.G., Oren, R., Manzoni, S., Higgins, C., Parlange, M.B., 2012. Evapotranspiration: A process driving mass transport and energy exchange in the soil-plant-atmosphere-climate system. Rev. Geophys. 50 (3) https://doi.org/10.1029/2011RG000366.
Khan, M.S., Liaqat, U.W., Baik, J., Choi, M., 2018. Stand-alone uncertainty characterization of GLEAM, GLDAS and MOD16 evapotranspiration products using an extended triple collocation approach. Agric. For. Meteorol. 252, 256–268. https://doi.org/10.1016/j.agrformet.2018.01.022.
Koch, J., Jensen, K.H., Stisen, S., 2015. Toward a true spatial model evaluation in distributed hydrological modeling: Kappa statistics, Fuzzy theory, and EOF-analysis benchmarked by the human perception and evaluated against a modeling case study. Water Resour. Res. 51 (2), 1225–1246. https://doi.org/10.1002/2014WR016607.
Koch, J., Mendiguren, G., Mariethoz, G., Stisen, S., 2017. Spatial sensitivity analysis of simulated land surface patterns in a catchment model using a set of innovative spatial performance metrics. J. Hydrometeorol. 18, 1121–1142. https://doi.org/10.1175/JHM-D-16-0148.1.
Koch, J., Siemann, A., Stisen, S., Sheffield, J., 2016. Spatial validation of large-scale land surface models against monthly land surface temperature patterns using innovative performance metrics. J. Geophys. Res. 121 (10), 5430–5452. https://doi.org/10.1002/2015JD024482.
Koren, V., Schaake, J., Mitchell, K., Duan, Q.-Y., Chen, F., Baker, J.M., 1999. A parameterization of snowpack and frozen ground intended for NCEP weather and climate models. J. Geophys. Res. Atmos. 104 (D16), 19569–19585. https://doi.org/10.1029/1999JD900232.
Koster, R., Suarez, M., 1992. Modeling the land surface boundary in climate models as a composite of independent vegetation stands. J. Geophys. Res. Atmos. 97, 2697–2715. https://doi.org/10.1029/91JD01696.
Koster, R.D., Suarez, M.J., Ducharne, A., Stieglitz, M., Kumar, P., 2000. A catchment-based approach to modeling land surface processes in a general circulation model 1 Model structure. J. Geophys. Res. Atmos. 105 (D20), 24809–24822. https://doi.org/10.1029/2000JD900327.
Kumar, S.V., Holmes, T.R., Mocko, D.M., Wang, S., Peters-Lidard, C.D., 2018. Attribution of flux partitioning variations between land surface models over the continental U.S. Remote Sens. 10, 751. https://doi.org/10.3390/rs10050751.
Kumar, S.V., Peters-Lidard, C.D., Tian, Y., Houser, P.R., Geiger, J., Olden, S., Lighty, L., Eastman, J.L., Doty, B., Dirmeyer, P., Mitchell, J.K., Wood, E.F., Sheffield, J., 2006. Land information system – An interoperable framework for high resolution land surface modeling. Environ. Modell. Softw. 21, 1402–1415. https://doi.org/10.1016/j.envsoft.2005.07.004.
Kumar, S.V., Wang, S., Mocko, D.M., Peters-Lidard, C.D., Xia, Y., 2017. Similarity assessment of land surface model outputs in the North American land data assimilation system (NLDAS). Water Resour. Res. 53 (11), 8941–8965. https://doi.org/10.1002/2017WR020635.
Lawston, P.M., Santanello, J.A., Franz, T.E., Rodell, M., 2017. Assessment of irrigation physics in a land surface modelling framework using non-traditional and human practice datasets. Hydrol. Earth Syst. Sci. 21, 2953–2966. https://doi.org/10.5194/hess-21-2953-2017.
Li, X., Liu, S., Li, H., Ma, Y., Wang, J., Zhang, Y., Xu, Z., Xu, T., Song, L., Yang, X., Lu, Z., Wang, Z., Guo, Z., 2018. Intercomparison of six upscaling evapotranspiration methods: From site to the satellite pixel. J. Geophys. Res. Atmos. 123 (13), 6777–6803. https://doi.org/10.1029/2018JD028422.
Liang, X., Lettenmaier, D.P., Wood, E.F., Burges, S.J., 1994. A simple hydrologically based model of land surface water and energy fluxes for general circulation models. J. Geophys. Res. Atmos. 99, 14415–14428. https://doi.org/10.1029/94JD00483.
Liu, S.M., Xu, Z.W., Wang, W.Z., Jia, Z.Z., Zhu, M.J., Bai, J., Wang, J.M., 2011. A comparison of eddy–covariance and large aperture scintillometer measurements with respect to the energy balance closure problem. Hydrol. Earth Syst. Sci. 15 (4), 1291–1306. https://doi.org/10.5194/hess-15-1291-2011.
Liu, W., Wang, L., Zhou, J., Li, Y., Sun, F., Fu, G., Li, X., Sang, Y.F., 2016. A worldwide evaluation of basin-scale evapotranspiration estimates against the water balance method. J. Hydrol. 538, 82–95. https://doi.org/10.1016/j.jhydrol.2016.04.006.
Long, D.i., Longuevergne, L., Scanlon, B.R., 2014. Uncertainty in evapotranspiration from land surface modeling, remote sensing, and GRACE satellites. Water Resour. Res. 50 (2), 1131–1151. https://doi.org/10.1002/2013WR014581.
Ma, N., Szilagyi, J., Zhang, Y., Liu, W., 2019. Complementary-relationship-based modeling of terrestrial evapotranspiration across china during 1982–2012: Validations and Spatiotemporal Analyses. J. Geophys. Res. Atmos. 124 (8), 4326–4351. https://doi.org/10.1029/2018JD029850.
Mao, J., Fu, W., Shi, X., Ricciuto, D.M., Fisher, J.B., Dickinson, R.E., Wei, Y., Shem, W., Piao, S., Wang, K., Schwalm, C.R., Tian, H., Mu, M., Arain, A., Ciais, P., Cook, R., Dai, Y., Hayes, D., Hoffman, F.M., Huang, M., Huang, S., Huntzinger, D.N., Ito, A., Jain, A., King, A.W., Lei, H., Lu, C., Michalak, A.M., Parazoo, N., Peng, C., Peng, S., Poulter, B., Schaefer, K., Jafarov, E., Thornton, P.E., Wang, W., Zeng, N., Zeng, Z., Zhao, F., Zhu, Q., Zhu, Z., 2015. Disentangling climatic and anthropogenic controls on global terrestrial evapotranspiration trends. Environ. Res. Lett. 10 (9), 094008. https://doi.org/10.1088/1748-9326/10/9/094008.
Martens, B., Miralles, D.G., Lievens, H., van der Schalie, R., de Jeu, R.A.M., Fernández-Prieto, D., Beck, H.E., Dorigo, W.A., Verhoest, N.E.C., 2017. GLEAM v3: satellite-based land evaporation and root-zone soil moisture. Geosci. Model Dev. 10, 1903–1925. https://doi.org/10.5194/gmd-10-1903-2017.
Mendiguren, G., Koch, J., Stisen, S., 2017. Spatial pattern evaluation of a calibrated national hydrological model – A remote-sensing-based diagnostic approach. Hydrol. Earth Syst. Sci. 21, 5987–6005. https://doi.org/10.5194/hess-21-5987-2017.
Miralles, D.G., Teuling, A.J., van Heerwaarden, C.C., Vilà-Guerau de Arellano, J., 2014. Mega-heatwave temperatures due to combined soil desiccation and atmospheric heat accumulation. Nat. Geosci. 7 (5), 345–349. https://doi.org/10.1038/ngeo2141.
Mitchell, K.E., Lohmann, D., Houser, P.R., Wood, E.F., Schaake, J.C., Robock, A., Cosgrove, B.A., Sheffield, J., Duan, Q., Luo, L., Higgins, R.W., Pinker, R.T., Tarpley, J.D., Lettenmaier, D.P., Marshall, C.H., Entin, J.K., Pan, M., Shi, W., Koren, V., Meng, J., Ramsay, B.H., Bailey, A.A., 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. Atmos. 109, D07S90. https://doi.org/10.1029/2003JD003823.
Monteith, J.L., 1965. Evaporation and environment. Symp. Soc. Exp. Biol. 19, 205–224.
Mu, Q., Heinsch, F.A., Zhao, M., Running, S.W., 2007. Development of a global evapotranspiration algorithm based on MODIS and global meteorology data. Remote Sens. Environ. 111 (4), 519–536. https://doi.org/10.1016/j.rse.2007.04.015.
Mu, Q., Zhao, M., Running, S.W., 2011. Improvements to a MODIS global terrestrial evapotranspiration algorithm. Remote Sens. Environ. 115 (8), 1781–1800. https://doi.org/10.1016/j.rse.2011.02.019.
Mueller, B., Seneviratne, S.I., Jimenez, C., Corti, T., Hirschi, M., Balsamo, G., Ciais, P., Dirmeyer, P., Fisher, J.B., Guo, Z., Jung, M., Maignan, F., McCabe, M.F., Reichle, R., Reichstein, M., Rodell, M., Sheffield, J., Teuling, A.J., Wang, K., Wood, E.F., Zhang, Y., 2011. Evaluation of global observations-based evapotranspiration datasets and IPCC AR4 simulations. Geophys. Res. Lett. 38 (6) https://doi.org/10.1029/2010GL046230.
Oki, T., Kanae, S., 2006. Global hydrological cycles and world water resources. Science 313, 1068–1072. https://doi.org/10.1126/science.1128845.
Peters-Lidard, C.D., Kumar, S.V., Mocko, D.M., Tian, Y., 2011. Estimating evapotranspiration with land data assimilation systems. Hydrol. Process. 25, 3979–3992. https://doi.org/10.1002/hyp.8387.
Refsgaard, J.C., 2001. Towards a formal approach to calibration and validation of models using spatial data. In: Grayson, R., Blöschl, G. (Eds.), Spatial Patterns in Catchment Hydrology: Observations and Modelling. Cambrige Univ. Press, Cambrige, pp. 329–354.
Renard, P., Allard, D., 2013. Connectivity metrics for subsurface flow and transport. Adv. Water Resour. 51, 168–196. https://doi.org/10.1016/j.advwatres.2011.12.001.
Reynolds, C.A., Jackson, T.J., Rawls, W.J., 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 (12), 3653–3662. https://doi.org/10.1029/2000WR900130.
Roberts, N.M., Lean, H.W., 2008. Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. Mon. Wea. Rev. 136, 78–97. https://doi.org/10.1175/2007MWR2123.1.
Rodell, M., Houser, P.R., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C.-J., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., Entin, J.K., Walker, J.P., Lohmann, D., Toll, D., 2004. The global land data assimilation system. Bull. Am. Meteorol. Soc. 85 (3), 381–394. https://doi.org/10.1175/BAMS-85-3-381.
Sheffield, J., Wood, E.F., Roderick, M.L., 2012. Little change in global drought over the past 60 years. Nature. 491 (7424), 435–438. https://doi.org/10.1038/nature11575.
Srivastava, A., Deb, P., Kumari, N., 2020. Multi-model approach to assess the dynamics of hydrologic components in a tropical ecosystem. Water. Resour. Manag. 34 (1), 327–341. https://doi.org/10.1007/s11269-019-02452-z.
Sun, R., Yuan, H., Liu, X., Jiang, X., 2016. Evaluation of the latest satellite-gauge precipitation products and their hydrologic applications over the Huaihe River basin. J. Hydrol. 536, 302–319. https://doi.org/10.1016/j.jhydrol.2016.02.054.
Sun, R., Hernández, F., Liang, X., Yuan, H., 2020. A calibration framework for high-resolution hydrological models using a multiresolution and heterogeneous strategy. Water Resour. Res. 56 https://doi.org/10.1029/2019WR026541 e2019WR026541.

Sun, S., Chen, B., Shao, Q., Chen, J., Liu, J., Zhang, X.J., Zhang, H., Lin, X., 2017. Modeling evapotranspiration over China's landmass from 1979 to 2012 using multiple land surface models: Evaluations and analyses. J. Hydrometeorol. 18, 1185–1203. https://doi.org/10.1175/JHM-D-16-0212.1.

Szilagyi, J., Crago, R., Qualls, R., 2017. A calibration-free formulation of the complementary relationship of evaporation for continental-scale hydrology. J. Geophys. Res. Atmos. 122 (1), 264–278. https://doi.org/10.1002/2016JD025611.

Vicente-Serrano, S.M., Miralles, D.G., Domínguez-Castro, F., Azorin-Molina, C., El Kenawy, A., McVicar, T.R., Tomás-Burguera, M., Beguería, S., Maneta, M., Peña-Gallardo, M., 2018. Global assessment of the standardized evapotranspiration deficit index (SEDI) for drought analysis and monitoring. J. Clim. 31 (14), 5371–5393. https://doi.org/10.1175/JCLI-D-17-0775.1.

Wang, K., Dickinson, R.E., 2012. A review of global terrestrial evapotranspiration: Observation, modeling, climatology, and climatic variability. Rev. Geophys. 50 (2) https://doi.org/10.1029/2011RG000373.

Wang, S., Pan, M., Mu, Q., Shi, X., Mao, J., Brümmer, C., Jassal, R.S., Krishnan, P., Li, J., Andrew Black, T., 2015. Comparing evapotranspiration from eddy covariance measurements, water budgets, remote sensing, and land surface models over Canada. J. Hydrometeorol. 16, 1540–1560. https://doi.org/10.1175/JHM-D-14-0189.1.

Wang, W., Cui, W., Wang, X., Chen, X., 2016. Evaluation of GLDAS-1 and GLDAS-2 forcing data and Noah model simulations over China at the monthly scale. J. Hydrometeorol. 17, 2815–2833. https://doi.org/10.1175/JHM-D-15-0191.1.

Wealands, S.R., Grayson, R.B., Walker, J.P., 2005. Quantitative comparison of spatial fields for hydrological model assessment—some promising approaches. Adv. Water Resour. 28, 15–32. https://doi.org/10.1016/j.advwatres.2004.10.001.

Western, A.W., Blöschl, G., Grayson, R.B., 2001. Toward capturing hydrologically significant connectivity in spatial patterns. Water Resour. Res. 37 (1), 83–97. https://doi.org/10.1029/2000WR900241.

Wilson, K., Goldstein, A., Falge, E., Aubinet, M., Baldocchi, D., Berbigier, P., Bernhofer, C., Ceulemans, R., Dolman, H., Field, C., Grelle, A., Ibrom, A., Law, B.E., Kowalski, A., Meyers, T., Moncrieff, J., Monson, R., Oechel, W., Tenhunen, J., Valentini, R., Verma, S., 2002. Energy balance closure at FLUXNET sites. Agric. For. Meteorol. 113 (1-4), 223–243. https://doi.org/10.1016/S0168-1923(02)00109-0.

Xia, Y., Hobbins, M.T., Mu, Q., Ek, M.B., 2015. Evaluation of NLDAS-2 evapotranspiration against tower flux site observations. Hydrol. Process. 29 (7), 1757–1771. https://doi.org/10.1002/hyp.v29.710.1002/hyp.10299.

Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., Luo, L., Alonge, C., Wei, H., Meng, J., Livneh, B., Lettenmaier, D., Koren, V., Duan, Q., Mo, K., Fan, Y., Mocko, D., 2012. Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. J. Geophys. Res. Atmos. 117 https://doi.org/10.1029/2011JD016048.

Xia, Y., Mocko, D., Huang, M., Li, B., Rodell, M., Mitchell, K.E., Cai, X., Ek, M.B., 2017. Comparison and assessment of three advanced land surface models in simulating terrestrial water storage components over the United States. J. Hydrometeorol. 18, 625–649. https://doi.org/10.1175/JHM-D-16-0112.1.

Xia, Y., Mocko, D.M., Wang, S., Pan, M., Kumar, S.V., Peters-Lidard, C.D., Wei, H., Wang, D., Eka, M.B., 2018. Comprehensive evaluation of the Variable Infiltration Capacity (VIC) model in the North American Land Data Assimilation System. J. Hydrometeorol. 19, 1853–1879. https://doi.org/10.1175/JHM-D-18-0139.1.

Xu, T., Guo, Z., Liu, S., He, X., Meng, Y., Xu, Z., Xia, Y., Xiao, J., Zhang, Y., Ma, Y., Song, L., 2018. Evaluating different machine learning methods for upscaling evapotranspiration from flux towers to the regional scale. J. Geophys. Res. Atmos. 123 (16), 8674–8690. https://doi.org/10.1029/2018JD028447.

Xu, T., Guo, Z., Xia, Y., Ferreira, V.G., Liu, S., Wang, K., Yao, Y., Zhang, X., Zhao, C., 2019. Evaluation of twelve evapotranspiration products from machine learning, remote sensing and land surface models over conterminous United States. J. Hydrol. 578, 124105. https://doi.org/10.1016/j.jhydrol.2019.124105.

Xu, Z., Ma, Y., Liu, S., Shi, W., Wang, J., 2017. Assessment of the energy balance closure under advective conditions and its impact using remote sensing data. J. Appl. Meteorol. Climatol. 56, 127–140. https://doi.org/10.1175/JAMC-D-16-0096.1.

Yang, K., Zhu, L., Chen, Y., Zhao, L., Qin, J., Lu, H., Tang, W., Han, M., Ding, B., Fang, N., 2016. Land surface model calibration through microwave data assimilation for improving soil moisture simulations. J. Hydrol. 533, 266–276. https://doi.org/10.1016/j.jhydrol.2015.12.018.

Yu, G.-R., Wen, X.-F., Sun, X.-M., Tanner, B.D., Lee, X., Chen, J.-Y., 2006. Overview of ChinaFLUX and evaluation of its eddy covariance measurement. Agric. For. Meteorol. 137 (3-4), 125–137. https://doi.org/10.1016/j.agrformet.2006.02.011.

Zeng, Z., Wang, T., Zhou, F., Ciais, P., Mao, J., Shi, X., Piao, S., 2014. A worldwide analysis of spatiotemporal changes in water balance-based evapotranspiration from 1982 to 2009. J. Geophys. Res. Atmos. 119 (3), 1186–1202. https://doi.org/10.1002/2013JD020941.

Zhang, B., Xia, Y., Long, B., Hobbins, M., Zhao, X., Hain, C., Li, Y., Anderson, M.C., 2020. Evaluation and comparison of multiple evapotranspiration data models over the contiguous United States: Implications for the next phase of NLDAS (NLDAS-Testbed) development. Agric. For. Meteorol. 280, 107810. https://doi.org/10.1016/j.agrformet.2019.107810.

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

Zhang, Y., Kong, D., Gan, R., Chiew, F.H., McVicar, T.R., Zhang, Q., Yang, Y., 2019. Coupled estimation of 500 m and 8-day resolution global evapotranspiration and gross primary production in 2002–2017. Remote Sens. Environ. 222, 165–182. https://doi.org/10.1016/j.rse.2018.12.031.