Title
Evaluating three evapotranspiration estimates from model of different complexity over China using the ILAMB benchmarking system

Authors
Wu, Genan
Cai, Xitian
Keenan, Trevor F
et al.

Publication Date
2020-11-01

DOI
10.1016/j.jhydrol.2020.125553

Peer reviewed

Evaluating three evapotranspiration estimates from model of different complexity over China using the ILAMB benchmarking system

Genan Wu[1,2,3,4], Xitian Cai[3], Trevor F. Keenan[3,4], Shenggong Li[1,2,*], Xiangzhong Luo[3,4], Joshua B. Fisher[5], Ruochen Cao[6], Fa Li[3,7], Adam J Purdy[5], Wei Zhao[1,6], Xiaomin Sun[1,2], Zhongmin Hu[6,8,*]

1 Synthesis Research Center of Chinese Ecosystem Research Network, Key Laboratory of Ecosystem Network Observation and Modeling, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China

2 College of Resources and Environment, University of Chinese Academy of Sciences, Beijing, China

3 Lawrence Berkeley National Laboratory, Berkeley, CA, USA

4 UC Berkeley, Berkeley, CA, USA

5 Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA

6 School of Geography, South China Normal University, Guangzhou, China

7 State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, China

8 Southern Marine Science and Engineering Guangdong Laboratory, Zhuhai, China.

*Corresponding author 1 : Zhongmin Hu, Email address : huzm@m.scnu.edu.cn ;Tel./fax : +86 10 64889039.

*Corresponding author 2 : Shenggong Li, Email address : lisg@igsnrr.ac.cn.; Tel./fax : +86 10 64889039.

Abstract: Land surface models range in complexity of terrestrial evapotranspiration, yet it is unknown how model complexity translates to accuracy of modeled evapotranspiration estimates. Here, we use the International Land Model Benchmarking system to assess ET estimates from three models of varying complexity driven by the same forcing datasets: an earth system model, a terrestrial biosphere model, and a stand-alone ET model. The performance assessment includes both temporal and spatial evaluation, and different plant functional types across China. Our results indicate that the most complex model, an earth system model, performed best against the benchmarking datasets and metrics. Terrestrial biosphere model performed best in simulating inter-annual variability of ET, while earth system model performed best in simulating the seasonal cycle. The more complex models (earth system model and terrestrial biosphere model) perform better in forest, shrub and crop ecosystems, while the simpler model (stand-alone ET model) perform better in grass ecosystems. Our study demonstrates the impact of model complexity on ET estimates and highlights directions for future ET model improvements.

Key words: Benchmarking, evapotranspiration model, model complexity

1. Introduction

Evapotranspiration (ET) is a key component of the global water budget and is crucial to agriculture and water management, the sustainability of ecosystems, and the water and carbon exchanges between land and atmosphere (Fisher et al., 2017). However, the estimation of large-scale ET from ground-based measurements alone remains challenging due to the sparse network of point observations and the high spatial and temporal variability of ET (Lu et al., 2017). To address this limitation, various terrestrial ET models have been developed (Jiménez et al., 2011; McCabe et al., 2016; Mueller et al., 2011; Vinukollu et al., 2011).

Terrestrial ET models play a vital role in diagnosing and predicting global water fluxes and in evaluating the impacts of changing climate (Mao et al., 2015). In recent years, a variety of physical process models have been developed to estimate the spatial distribution of evapotranspiration (ET) at various scales ranging from the stand scale to global. From empirical and semi-empirical method (i.e. Jackson model, Priestley-Taylor model) to physical processed method (i.e. Shuttleworth-Wallace model, Community Land Model), much progress has been made incorporate more physical processes into ET simulations (Bonan et al., 2013; Jackson, 1985; Priestley and Taylor, 1972; Shuttleworth and Wallace, 1985). In addition, some statistic and machine learning methods were used to improve ET models performance and accuracy (Adnan et al., 2020; Alizamir et al., 2020). As ET models become increasingly complex and the number of model parameters rapidly expands, there is a growing need for a comprehensive and multifaceted evaluation of the performance of models of different levels of complexity (Haughton et al., 2016; Hogue et al., 2006). In this study, "complexity" is defined in terms of the number of process-related variables and parameters and the hierarchy of model structure. In terrestrial ET models, for example, the Priestley-Taylor model (Priestley and Taylor, 1972)—a simplification of the Penman-Monteith equation (Monteith, 1965)—requires less forcing data and thus does not consider explicitly the impact of vapor pressure deficit (VPD) or canopy resistance. This method is convenient to use in the absence of detailed meteorological measurements. By contrast, the Penman-Monteith model and the Shuttleworth-Wallace model (Shuttleworth and Wallace, 1985) consider complex biogeochemical and biogeophysical land surface processes and therefore require more meteorological measurements and parameters (Fisher et al., 2011). Specifically, the Shuttleworth-Wallace model partitions ET into soil water evaporation and plant transpiration and contains more complexity estimation of ET processes.

In recent decades, earth system models (ESM) which simulate biogeochemical processes on the land surface, which are fully coupled with physical climate simulations, have been developed rapidly and widely used (Bonan and Doney, 2018). Meanwhile, the estimation of the physical-process variables of an ESM such as ET is becoming increasingly comprehensive and sophisticated. Compared to other terrestrial ET models, ESM require higher temporal-spatial resolution forcing data and physical parameters (Mueller et al., 2013). Although more complicated ET models can provide more details involved in atmosphere-terrestrial water exchange, they are also potentially prone to greater uncertainties propagated from other related processes (Orth et al., 2015). There remains a lack of knowledge on the optimal complexity of ET models on the regional scale.

Model benchmarking has emerged as an effective approach to evaluate model performance relative to multiple observational constraints as well as other models (Collier et al. 2018). Most recently, the International Land Model Benchmarking (ILAMB) System (Collier et al., 2018; Luo et al., 2012; Stofferahn et al., 2019), the ESM Evaluation Tool (Eyring et al., 2016), the Program for Climate Model Diagnosis and Intercomparison Metrics Package (Gleckler et al., 2016) and other benchmarking system were created to explore land surface model intercomparison and facilitate internationally accepted benchmarks (Schwalm et al., 2013).

The aim of this paper is to leverage the ILAMB benchmarking tool to assess the performance among three terrestrial ET models with various levels of complexity at the regional scale (Polhamus et al., 2013). Taking China as an example research area, these objectives are accomplished by evaluating the performance of three ET models of varying levels of complexity for: 1) inter-annual and seasonal variation; 2) spatial variation; and, 3) different plant functional types (PFT). To facilitate the comparison, we used the same forcing datasets for each of the three ET models, in order to limit the uncertainty of the forcing data (Badgley et al., 2015) and focus on the effect of model complexity.

2. Methodology

2.1 ILAMB Description

As land surface models become increasingly complex and observational data volumes rapidly expand, there is a growing need for comprehensive and multifaceted evaluation of model fidelity. Building on past model evaluation work (Randerson et al., 2009), Luo et al. (2012) and Collier et al. (2018) developed an extensible model benchmarking package in support of the goals of the International Land Model Benchmarking (ILAMB) activity. The ILAMB benchmarking system compares model estimates against the best-available observations and observation-based extrapolations, including atmosphere CO_2 concentrations, surface fluxes, hydrology, soil carbon and nutrient biogeochemistry, ecosystem processes and states, and vegetation dynamics.

To evaluate the differences between reference and model datasets, a variety of statistical approaches have been adopted, including calculations of bias, root-mean-square error (RMSE), phase, amplitude, spatial distribution, Taylor diagrams and scores, functional relationship metrics, and perturbation and sensitivity tests. Bias is calculated as follows:

$$bias(\mathbf{x}) = \overline{v_{\mathrm{mod}}}(\mathbf{x}) - \overline{v_{\mathrm{ref}}}(\mathbf{x}) \qquad (1)$$

The variable **x** is spatial domain which represents the areas created by cell boundaries or the areas connected with data sites. $\overline{v_{\mathrm{mod}}}(\mathbf{x})$ is the mean value over time of a modelled dataset. $\overline{v_{\mathrm{ref}}}(\mathbf{x})$ is the mean value over time of a reference dataset. We then nondimensionalized the biases into a relative error using the centralized RMS (Root Mean Square) of the reference dataset following equation (2):

$$\mathrm{crms}(x) = \sqrt{\frac{1}{\mathrm{t_f}-\mathrm{t_0}}\int_{\mathrm{t_0}}^{\mathrm{t_f}}(\mathrm{v_{ref}}(t,x) - \overline{\mathrm{v_{ref}}}(x))^2\ \mathrm{d}t} \qquad (2)$$

The variable t is the temporal domain which is defined by the beginning and end of studied period. The relative error in bias is:

$$\varepsilon_{\mathrm{bias}}(x) = |\mathrm{bias}(x)|/\mathrm{crms}(x) \tag{3}$$

The bias score as a function of space is:

$$\mathrm{s}_{\mathrm{bias}}(x) = \mathrm{e}^{-\varepsilon_{\mathrm{bias}}(x)} \tag{4}$$

And the scalar score

$$\mathrm{S}_{\mathrm{bias}} = \overline{\overline{\mathrm{s}_{\mathrm{bias}}}}(x) \tag{5}$$

that is, the spatially integrated bias score. RMSE over the period of the reference dataset is estimated as follows:

$$\mathrm{RMSE}(x) = \sqrt{\frac{1}{\mathrm{t_f}-\mathrm{t_0}}\int_{\mathrm{t_0}}^{\mathrm{t_f}}(\mathrm{v_{mod}}(t,x) - \mathrm{v_{ref}}(t,x))^2 \ \mathrm{d}t} \tag{6}$$

To score the RMSE, we use the methods similar to Eq. (2-5). Please refer to Collier et al. (2018) for more details. ILAMB evaluates the phase shift of the annual cycle of data sets that have intra-annual variability by comparing the timing of the maximum value in a year, $c(v)$ within each. Then, we approximate the phase shift from the reference to model data sets by subtracting their respective $c(v)$,

$$\theta\ (x) = \arg\max_{\mathrm{t}}\left(\mathrm{c_{mod}}(t,x)\right) - \arg\max_{\mathrm{t}}\left(\mathrm{c_{ref}}(t,x)\right) \tag{7}$$

As the units for phase shift are consistent across all variables, no normalization is needed and we can remap the shift to the unit interval by

$$\mathrm{s_{phase}}(x) = \frac{1}{2}(1 + \cos(\frac{2\ \pi\ \ \theta\ (x)}{365})) \tag{8}$$

And the scalar score is:

$$\mathrm{S_{phase}} = \overline{\overline{\mathrm{s_{phase}}}}(x) \tag{9}$$

The score for the inter-annual variability is calculated by removing the annual cycle from both the reference and the model,

$$\mathrm{iav}_{\mathrm{ref}}(x) = \sqrt{\frac{1}{\mathrm{t_f}-\mathrm{t_0}}\int_{\mathrm{t_0}}^{\mathrm{t_f}}(\mathrm{v}_{\mathrm{ref}}(t,x)-\mathrm{c}_{\mathrm{ref}}(t,x))^2\ \mathrm{dt}} \tag{10}$$

$$\mathrm{iav}_{\mathrm{mod}}(x) = \sqrt{\frac{1}{\mathrm{t_f}-\mathrm{t_0}}\int_{\mathrm{t_0}}^{\mathrm{t_f}}(\mathrm{v}_{\mathrm{mod}}(t,x)-\mathrm{c}_{\mathrm{mod}}(t,x))^2\ \mathrm{dt}} \tag{11}$$

$$\varepsilon_{\mathrm{iav}}(x) = (\mathrm{iav}_{\mathrm{mod}}(\mathrm{x}) - \mathrm{iav}_{\mathrm{ref}}(x))/\mathrm{iav}_{\mathrm{ref}}(x) \tag{12}$$

and then computing a score as a function of space,

$$\mathrm{s}_{\mathrm{iav}}(x) = \mathrm{e}^{-\varepsilon_{\mathrm{iav}}(x)} \tag{13}$$

The scalar score is estimated by:

$$\mathrm{S}_{\mathrm{iav}} = \overline{\overline{\mathrm{s}_{\mathrm{iav}}}}(x) \tag{14}$$

To score the spatial distribution of the time averaged variable by generating a Taylor diagram (Taylor, 2001), we estimate the normalized standard deviation,

$$\sigma = \frac{\mathrm{stdev}(\overline{\mathrm{v}_{\mathrm{mod}}}(x))}{\mathrm{stdev}(\overline{\mathrm{v}_{\mathrm{ref}}}(x))} \tag{15}$$

and the spatial correlation R of the period mean values $\overline{v_{\mathrm{mod}}}(\mathbf{x})$ and $\overline{v_{\mathrm{ref}}}(\mathbf{x})$, and then assigning a score by the following relationship

$$\mathrm{S}_{\mathrm{dist}} = \frac{2(1+R)}{(\sigma+\frac{1}{\sigma})^2} \tag{16}$$

Where the main idea is that we penalize the sore when R and σ deviate from a value of 1. The overall score for a given variable and data product is a composite of the suite of metrics defined above. We use a weighted sum,

$$\mathrm{S}_{\mathrm{overall}} = \frac{\mathrm{S}_{\mathrm{bias}}+2\mathrm{S}_{\mathrm{rmse}}+\mathrm{S}_{\mathrm{phase}}+\mathrm{S}_{\mathrm{iav}}+\mathrm{S}_{\mathrm{dist}}}{1+2+1+1+1} \tag{17}$$

Where the RMSE score is doubled to emphasize its importance. In addition, we show the relative score (i.e., Z score), indicating which models or model versions perform better with respect to others contained in the overall analysis. More details of the underlying metrics are available in Collier et al. (2018).

2.2 Data Sets

To quantify and explain uncertainties and scale mismatches between reference datasets and model datasets, the ILAMB system developed a two-element rubric to weight each dataset (Table 1). The first weight of the datasets indicates the presence of quantitative uncertainty in the measurements themselves. A second weight reflects spatial and temporal coverage of the datasets. The reference datasets in ILAMB include in-situ observations (FLUXNET data), observation-satellite-meteorological ensemble data (FLUXCOM), multi ET product ensemble data, and remotely sensed data. As the aim of the ILAMB system is to evaluate model performance at the regional and decadal scales, users can give more weight to global products which have longer time series. The weights are combined multiplicatively to assign a total weight to each dataset. The weight for a given variable is then normalized relative to the sum of the weights of all the datasets for that variable (Eq. (18)).

In this study, we used four datasets to benchmark ET: FLUXNET, FLUXCOM, DOLCE, and GLEAM. Note that the FLUXCOM product was not used in inter-annual variability evaluation because it is known to poorly represent inter-annual variability (Jung et al. 2018). We assign the certainty weight and the scale weight as 3 and 5, respectively, for both the FLUXCOM and GLEAM datasets according to Collier et al. (2018). In addition, we assign the same weight for the FLUXNET and DOLCE dataset in order to more objective assessment (Table 1). For example, the normalized total weight of the FLUXNET dataset for the ET variable is estimated as:

$$w^{ET}_{FLUXNET} = \frac{3\times5}{3\times5+3\times5+3\times5+3\times5} \approx 25\% \quad (18)$$

Table 1. References and weighting of evapotranspiration (ET) data sets used to blend the overall score.

Reference datasets	Certainty	Scale	Source
FLUXNET	3	5	Pastorello et al. (2017)
FLUXCOM	3	5	Jung et al. (2019)
DOLCE	3	5	Hobeichi et al. (2018)
GLEAM	3	5	Martens et al. (2018)

The in-situ data used in this study were obtained from 12 FLUXNET sites in China (Figure 1): the Changbaishan temperate broad-leaved mixed forest (CN-Cha), Changling grassland (CN-Cng), Dangxiong alpine meadow (CN-Dan), Dinghushan subtropical evergreen broad-leaved forests (CN-Din), Duolun grassland (CN-Du2), Haibei alpine shrub wetland (CN-Ha2), Haibei alpine meadow (CN-Ha2), Qianyanzhou evergreen needleleaf forests (CN-Qia), Siziwang Grazed grassland (CN-Sw2), Yucheng cropland (YC), NeiMeng temperate grassland (NM), Xishuangbanna evergreen broadleaf forest (XSBN). Eddy covariance flux data of the 12 sites were extracted from the Tier 1 Subset product (FLUXNET2015 Dataset), which was downloaded directly from the FLUXNET website (http://FLUXNET.fluxdata.org/) and from ChinaFLUX (http://www.chinaflux.org/). Detailed descriptions are available in Table 2.

To assess the performance among three levels of complexity terrestrial ET models in different plant functional types (PFT), we used vegetation classification data (Figure 1) provided by Environmental and Ecological Science Data Center for West China, National Natural Science Foundation of China (http://westdc.westgis.ac.cn). The datasets are based on the results of vegetation field investigation from 1949 to 2000, satellite images, soil data and meteorological data.

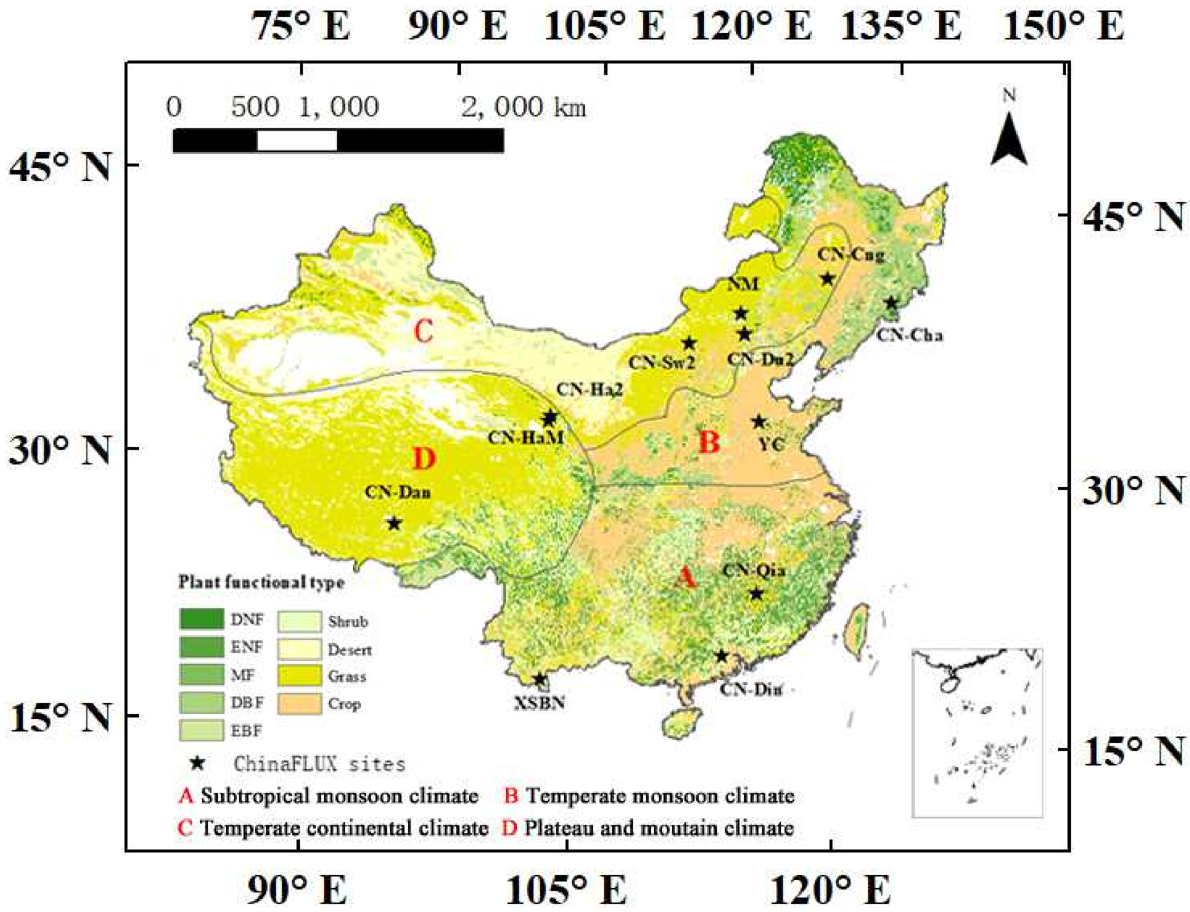

Figure 1. Locations of the 12 ChinaFLUX sites and distribution of plant functional type and climate zones.

Table 2. The list of ChinaFLUX sites used in this study.

Site ID	PFT	Lat (°N)	Lon(°W)	Data period	References

CN-Cha	MF	42.4	128.1	2005-2014	Guan et al. (2006)
CN-Cng	GRA	44.59	123.51	2007-2010	-
CN-Dan	GRA	30.50	91.07	2004-2008	Shi et al. (2006)
CN-Din	EBF	23.17	112.54	2003-2005	Zhang et al. (2010)
CN-Du2	GRA	42.05	116.28	2006-2008	Chen et al. (2009)
CN-Ha2	WET	37.61	101.33	2003-2005	-
CN-HaM	GRA	37.37	101.18	2002-2004	Kato et al. (2006)
CN-Qia	ENF	26.74	115.06	2003-2005	Yu et al. (2006)
CN-Sw2	GRA	41.79	111.9	2010-2012	-
YC	Crop	36.83	116.57	2003-2010	Yu et al. (2006)
NM	Grass	43.33	116.24	2004	Yu et al. (2006)
XSBN	EBF	21.93	101.27	2003-2010	Yu et al. (2006)

2.3 ET Model Descriptions

To limit the uncertainty of the forcing data and focus on the effect of different model complexity, we used the same meteorology datasets from 1980 to 2010 (GSWP3, https://www.isimip.org/gettingstarted/details/4/) and satellite remote sensing datasets (Normalized Difference Vegetation Index (NDVI) GIMMS product, https://glam1.gsfc.nasa.gov/) to run the three models. The simplest ET model is the Priestley Taylor-Jet Propulsion Laboratory (PT-JPL) model which is developed from Priestley-Taylor model (Fisher et al., 2008; Priestley and Taylor, 1972). The PT-JPL model incorporates a variety of data sources from meteorological data (i.e., net radiation (R_n), air temperature, vapor pressure) and satellite observations (NDVI, visible spectrum reflectance, near-infrared spectrum reflectance). We use the Shuttleworth-Wallace-Hu (SWH) model as a representative of intermediate complex models (Hu et al., 2013; Hu et al., 2017), which is developed based on the Shuttleworth-Wallace model and coupled light use efficiency model (Shuttleworth and Wallace, 1985). Meteorological data (i.e., air temperature, precipitation, relative humidity, wind speed, and R_n) and satellite products (i.e., NDVI) are the forcing data for the SWH model. We used the version 1 of the Energy Exascale Earth System Model (E3SM) Land Model (ELMv1) as a representative of the most complex ET model, which was branched from the version 4.5 of the Community Land Model (CLM4.5; Oleson et al. (2013)) with a specific version tag 4_5_71 (Cai et al., 2019). The forcing fields include surface air temperature, precipitation, wind speed, relative humidity, surface pressure, incoming solar radiation, and incoming longwave radiation. (Figure 2)

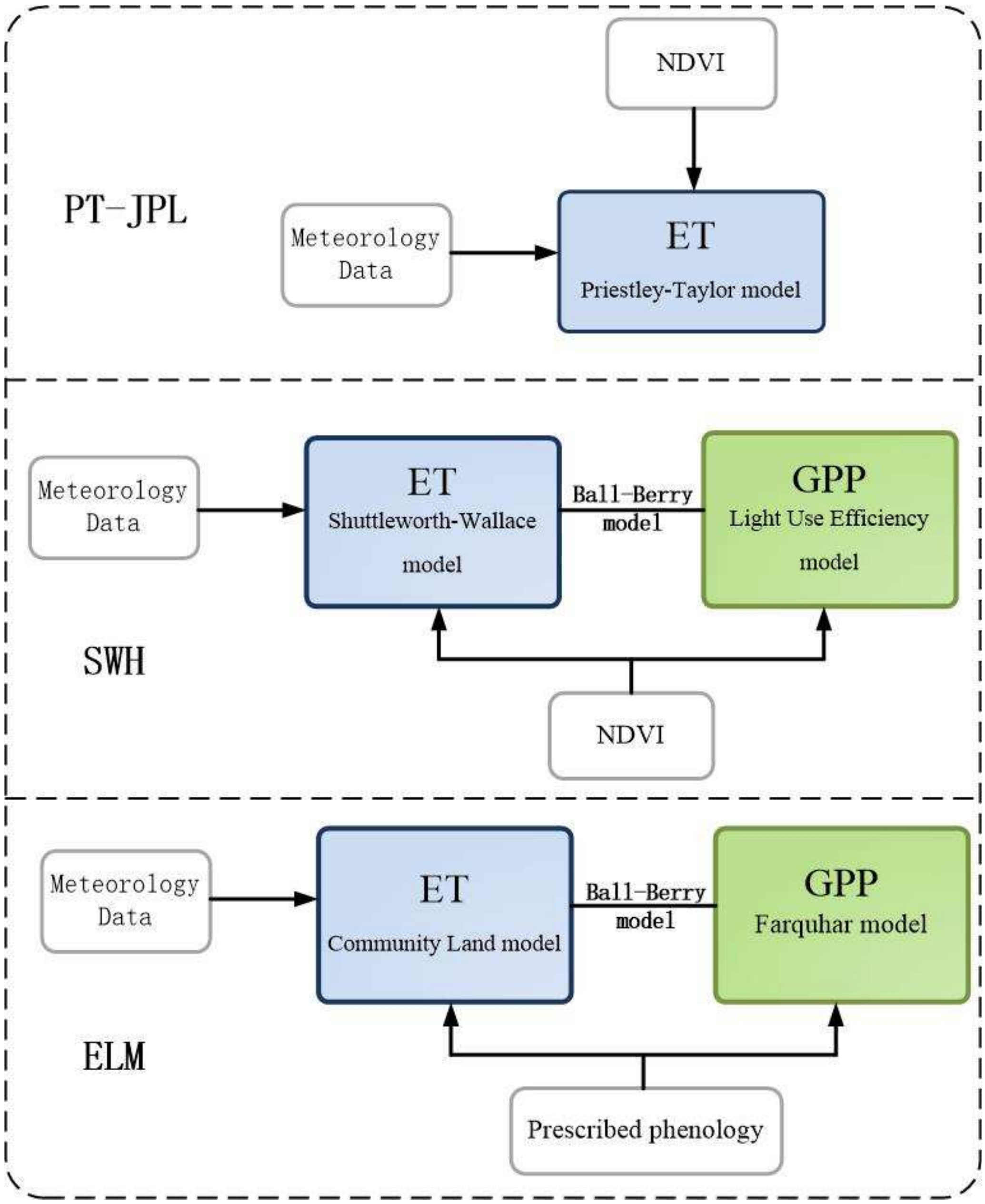

Figure 2. Evapotranspiration models: Priestley Taylor-Jet Propulsion Laboratory (PT-JPL) model, Shuttleworth-Wallace-Hu (SWH) model, and Energy Exascale Earth System Model Land Model (ELM).

3. Results

3.1 Overall performance

In ILAMB, compared with the reference datasets, we found a strong performance gradient among the three ET models. The most complicated model, ELM (overall absolute score: 0.71) perform best compared with reference datasets. The intermediate complexity model, with an overall score of SWH (0.67) is 0.04 lower than the ELM model. And the performance of the simplest model, PT-JPL (overall absolute score: 0.63) was lowest relative to the other models .

3.2 Inter-annual variability and seasonal cycle simulation performance

Compared with the inter-annual variability of reference ET dataset, the results (Figure 3) showed that 1) the simulation of inter-annual variability of the three ET models (ELM, SWH, PT-JPL) is better in eastern China than in western China; 2) the three ET models perform poor in some special geographical regions such as Qinghai-Tibet plateau and southwest mountains region; 3) the overall performance of inter-annual variability can be sorted in order of: SWH (mean score = 0.75) > ELM (mean score = 0.73) > PT-JPL (mean score = 0.70).

For the different climate region in China (Figure 3d), ELM model had the lowest score in simulating the inter-annual variability of ET in the plateau and mountain climate region (mean score = 0.47). There is a need to improve the ET inter-annual variability simulation of the three terrestrial ET models in the temperate continental climate region (mean score: ELM = 0.62, SWH = 0.68, JPL = 0.56). All three ET models perform equally well in the temperate monsoon climate region (mean score: ELM = 0.88, SWH = 0.89, JPL = 0.88). In the subtropical monsoon climate region, PT-JPL model had the worst performance of ET inter-annual variability simulation (mean score = 0.77).

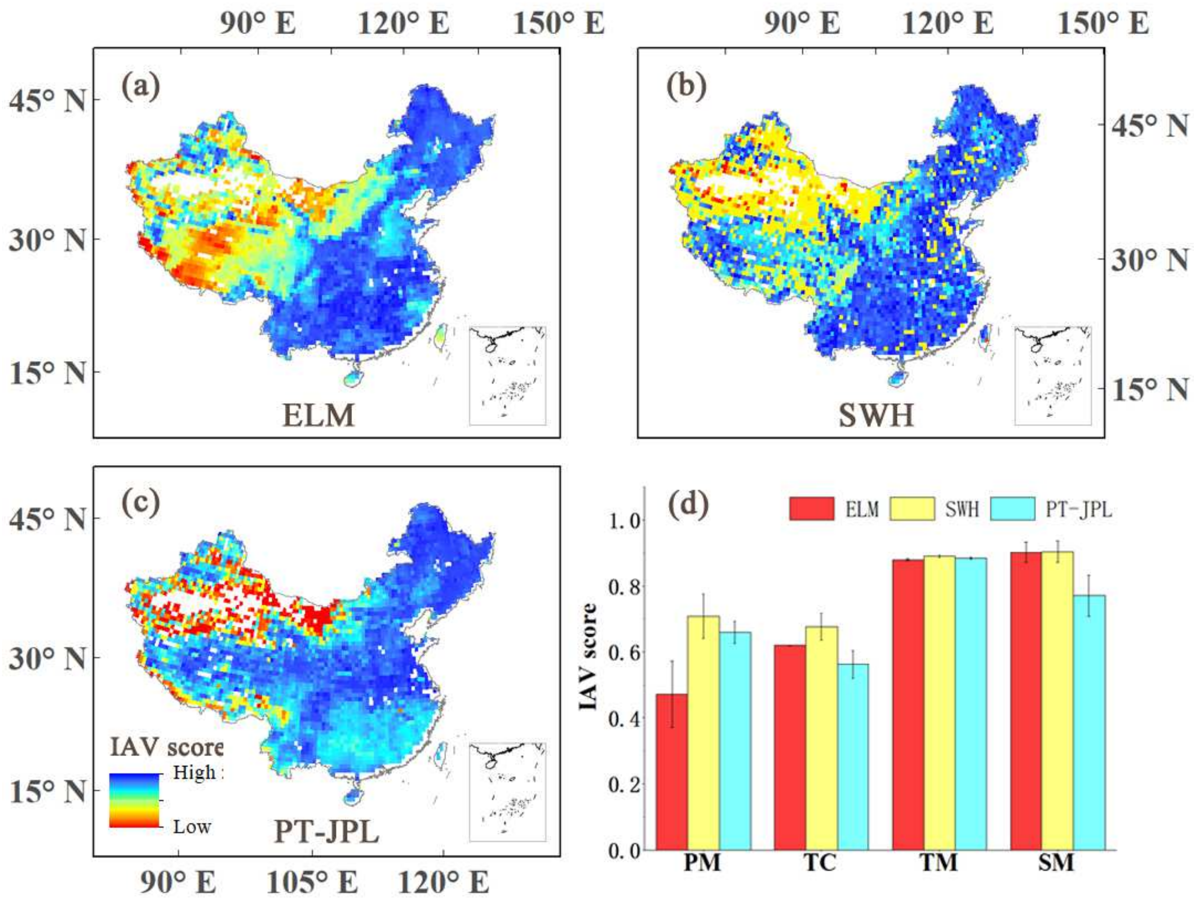

Figure 3. The spatial distribution of inter-annual variability (IAV) score of three models: (a) ELM, (b) SWH and (c) PT-JPL and (d) the inter-annual variability score in different climate change: plateau and mountain climate (PM), temperate continental climate (TC), temperate monsoon climate (TM), subtropical monsoon climate (SM).

In terms of seasonal cycle score, which compares the timing of the maximum ET of the annual cycle between reference dataset and model dataset, ELM and PT-JPL (mean score=0.91, 0.90) performs better than SWH model (mean score=0.78). In northwestern and southwestern of China, the simulation of seasonal cycle of the three ET models had lower scores especially the SWH model (Figure 4).

In different climate region of China (Figure 4d), the three ET models had the worst performance in temperate continental climate region especially SWH model (mean score: ELM = 0.86, SWH = 0.69, JPL = 0.84). In the monsoon climate region, the three ET models perform better than plateau and mountain climate region and temperate continental climate region. The ELM model performs well in different climate region of China.

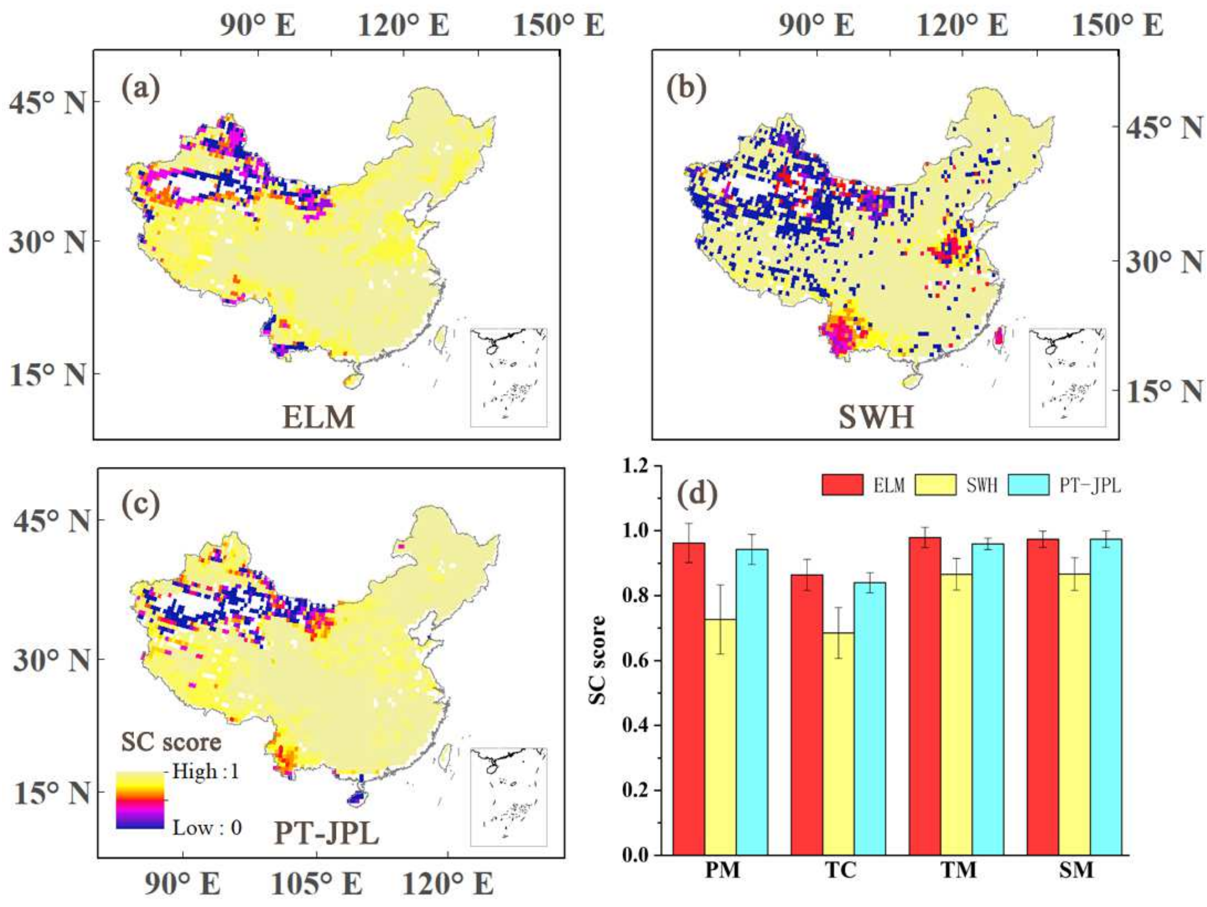

Figure 4. The spatial distribution of seasonal cycle (SC) score of three models: (a) ELM, (b) SWH and (c) PT-JPL and (d) the seasonal cycle score in different climate change.

3.3 Spatial variability performance

Taylor diagrams (Taylor, 2001) were used to analyze the spatial distribution of the time averaged ET. Taylor diagrams are particularly useful in evaluating multiple aspects of complex data series, since each graph shows a statistical summary of how well patterns match each other in terms of their correlation (r), their root mean square error (RMSE), and the normalized standard deviation (SD). The radial distance from the origin represents the amplitude of the ET variation (SD), normalized by the reference value (SD=1). The azimuthal angle of a particular point indicates its correlation to the reference. And the distance between a point and the reference shows the mean absolute difference between those datasets (RMSE). We used 31 year- averaged ET values of three models to assess spatial variability performance based on Taylor diagrams. As shown in Figure 5, the results indicated that 1) the correlation between ELM (r=0.96) and reference datasets is stronger than those of SWH (r=0.91) and PT-JPL (r=0.72); 2) even though the three model have different correlation, the standard deviation of three models has shown the similar distance relative to benchmark (SD_{ELM}=1.19, SD_{SWH}=0.81, $SD_{PT\text{-}JPL}$=1.20); 3) the ELM model has the smallest RMSE (0.32) when compared with SWH (0.41) and PT-JPL (0.79). On the whole, the most complex model, ELM which is closest to the benchmark has a good performance on spatial variability simulation.

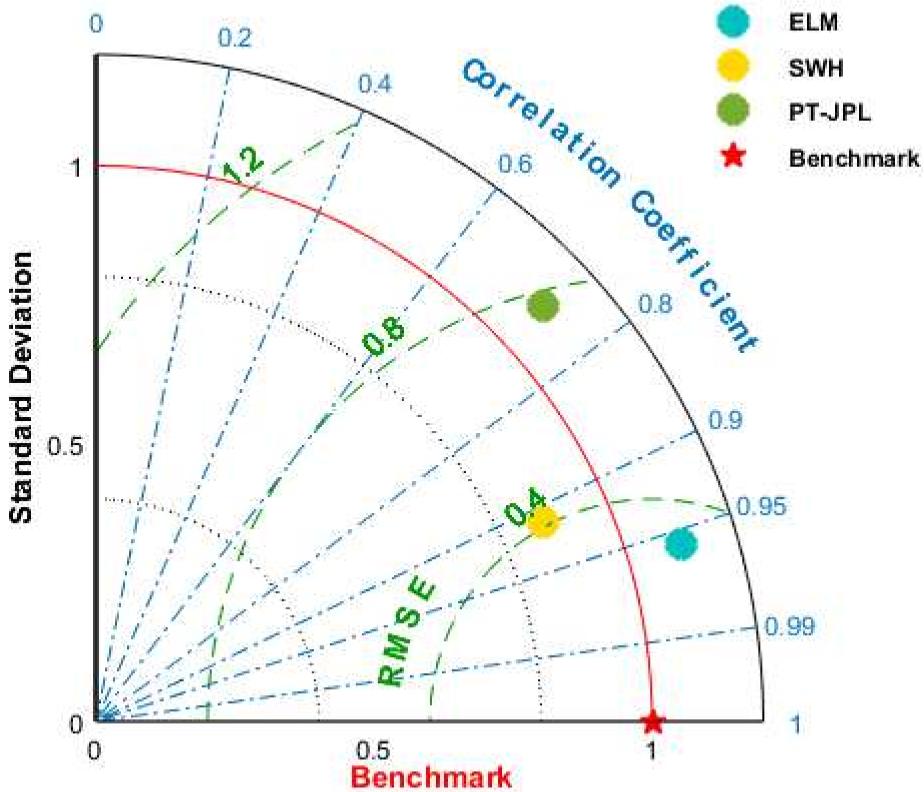

Figure 5. Taylor diagram showing correlation coefficient, RMSE, and standard deviation of spatial variability performance for the three ET models.

3.4 Model performance in different plant functional types

In different plant functional types (PFT), the three levels of complexity terrestrial ET models have different performance relative to the reference datasets. The most complicated ET model, ELM, shows the best performance in DNF, ENF, MF, DBF, and Crop (overall score = 0.75, 0.69, 0.70, 0.72, 0.71) but performs worst in Grass (overall score = 0.61). The best performance of the intermediate complexity model, SWH is achieved in EBF and Shrub (overall score = 0.72, 0.69). And the simplest model, PT-JPL have the best performance in Grass (overall score =0.71). Both of SWH and PT-JPL models has poor performance in forest ecosystems. Additionally, the relative score revealed that PT-JPL model perform worse in ENF, DBF, and EBF compared to the other models. (Figure 6)

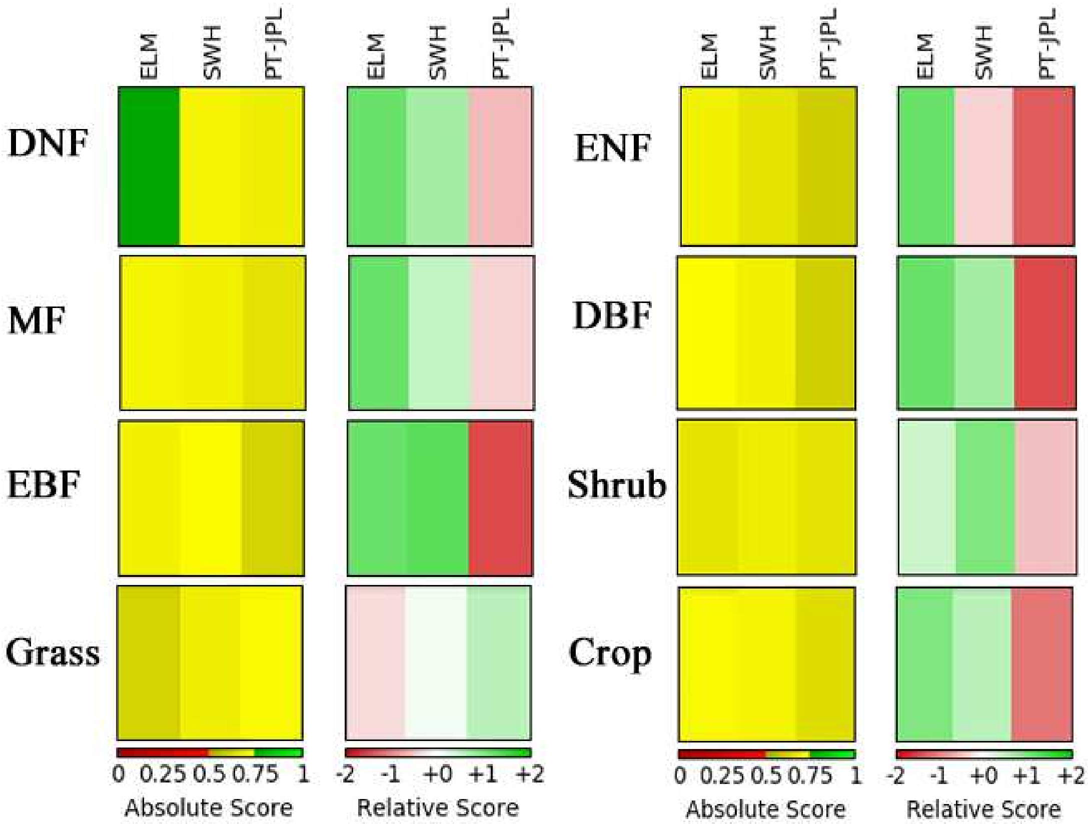

Figure 6. Overall score of ELM, SWH, PT-JPL model evapotranspiration estimates in different plant functional types

4. Discussion

4.1 Overall performance of the three levels of complexity terrestrial ET models

Our findings suggest that the performance of terrestrial ET models is related to some extent, but not entirely, to model complexity. The results showed that model complexity is positively correlated with ILAMB overall scores. As the ET models become increasingly complex, they contain an increasing number of biophysical, biochemical and biogeography descriptions. Several reports have shown that adding complexity to a land surface model may improve performance. Leplastrier (2002) investigated the performance of five modes of a land surface model, the Chameleon Surface Model (CHASM) and they found that the performance of more complex modes of CHASM is superior to more simple modes. Medici et al. (2012) analyzed three hydro-chemical models varying different level of complexity and the results presented that increased model complexity can improve performance if sufficient data are available for model testing. Our results support these earlier conclusions, though notable exceptions exist. However, there remains a lack of comparisons of different complexity ET models and exploration of the differences in their mechanisms. In future work on ET model evaluation, large ensembles of models of different complexity are needed in order to compare and improve ET modeling, in addition to the incorporation of more observed ET datasets as benchmark datasets in the ILAMB system.

4.2 Temporal and spatial simulation performance

Given that direct model evaluation is possible only with contemporary *in-situ* observations, it is difficult to assess the models' capacities to capture spatial variation at large scale. Khosa et al. (2019) evaluated and calibrated surface, empirical and satellite-based models performance including inter-annual variation and seasonal cycle performance compared with in situ ET measurement in South Africa. Ma et al. (2019a) validated a 31-year ET product by using plot-scale eddy covariance measurement and basin-scale water-balance-derived evapotranspiration rates and quantified the spatial and temporal variability of ET in China. However, we still lack a quantitative assessment of ET model performance distribution for inter annual variability and seasonal cycle. In this study, we leveraged the ILAMB system to enable improved testing of multiple terrestrial ET models, which used a wide variety of regional-scale gridded observations, site specific observations, and integrative observations to allow a more robust model benchmarking framework.

As shown in Figure 3, SWH performs best in terms of inter-annual variability simulation. And the simulation of inter-annual variability of the three ET models (ELM, SWH, PT-JPL) is poor in the northwest of China (temperate continental climate region). In the northwest arid region, temperature and precipitation experienced a sharp increasing in the past 50 years (Yang et al., 2018). The precipitation trend changed in 1987, and since then has been in a state of high volatility. Temperature experienced a “sharp” increase in 1997; since then, it has remained highly volatile, and the increasing trend slowed (Chen et al., 2015; Wang et al., 2017). Meanwhile，whether reanalysis climate product or interpolation climate data is effected by in situ measurements which is less distributed in the northwest of China. These may be one of the reasons for the poor inter-annual variability simulation performance in the northwest of China.

In some ecosystems that occupy particular eco-geographical locations and have special biogeochemical cycling, such as the Qinghai-Tibet Plateau (plateau and mountain climate region), the ELM model had the poorest performance for inter-annual variability. The atypical conditions in these regions could have affected the ELM soil thermal conductivity scheme (Farouki's scheme, Bonan et al. (2013)). Wang et al. (2014) found that the Farouki's scheme underestimated the upward shortwave radiation and overestimated the upward longwave and net radiation in Qinghai-Tibet Plateau. Several reports have shown that energy conditions are influential factors limiting ET in the entire Qinghai-Tibet Plateau especially at upper elevation (Ma et al., 2019b; Mingyue et al., 2019). Hence, reducing the uncertainty of soil thermal conductivity scheme may help improve the performance of the ET model in Qinghai-Tibet Plateau.

In terms of seasonal cycle simulation, ELM performed better than PT-JPL and the SWH model. In the northwest and southwest of China, the simulation of seasonal cycle of the three ET models had lower scores, especially SWH model. This is possibly due to the special geographical environment, in particular aridity of the northwest region and the southwest region (Yunnan Plateau). The lack of parameter localization for these regions is potentially responsible for the poor model performance.

In term of the spatial distribution simulation, ELM and SWH models have higher correlation coefficients with the reference dataset (0.96, 0.91, respectively), which is higher than the coefficient for PT-JPL model (0.72). On the other hand, ELM and the SWH model showed the smaller RMSE in comparison with the benchmark data. Considering the evidence above, we found that the more complex models (ELM, SWH) perform better for the ET spatial distribution than the simpler model (PT-JPL). A possible explanation for these results may be some key parameters of terrestrial ET model are space-time scale dependent and relate to traits in specific environmental (Chaney et al., 2016; Peaucelle et al., 2019). For the more complex models (ELM, SWH), the variations of key parameters are considered in the physical-process simulation in different PFT. It is therefore likely that the more complex models simulate spatial distribution better in China, due to their ability to better consider the variations and diveristy in the ecosystem characteristics.

4.3 Model performance in different plant functional types

The most complex ET model, ELM shows the best performance in most forest ecosystem (DNF, ENF, MF, DBF) and Crops. The best performance of the intermediate complexity model, SWH is achieved in EBF and Shrubs. And the simplest model, PT-JPL have the best performance in Grass.

ELM and SWH model coupled exchanges of energy, water, and carbon and incorporated photosynthesis process simulation. Plant stomata function as a controlling interface to regulate plant water loss and carbon dioxide uptake, and play a crucial role in ET and carbon exchange (Miner et al., 2017; Shan et al., 2019). Specifically, stomatal resistance is one of the largest drivers of ET under the situation that the canopy is fully coupled to the surrounding boundary layer, and therefore it provides links between ET and photosynthesis (De Kauwe et al., 2015; Shan et al., 2019). Both the ELM and SWH models incorporate Ball-Berry model

(Ball et al., 1987) to calculate stomatal resistance. SWH used a light use efficiency model (Running et al., 2004) to estimate the photosynthesis rate, which is a key parameter in the Ball-Berry model, while the photosynthesis rate in ELM is based on biochemical models (Collatz et al., 1992; Farquhar et al., 1980). ET integrates biochemical and biophysical land surface processes between the Earth's surface and atmosphere (Jung et al., 2010; Zhang et al., 2016). Coupling biochemical and biophysical processes in terrestrial ET models is thus expected to lead to improved performance. This improved process representation could explain why the ELM model performs better in particular in forest ecosystems, which have a more complex canopy structure.

Even though the PT-JPL model is developed using a semi-empirical satellite-based ET model, it performs best in grass ecosystems. This result may be explained by the fact that PT-JPL model performed better in water-limited regions, where remotely sensed information on dynamic vegetation responses to changes in water availability aid in the prediction of ET (Ershadi et al., 2014).

5. Conclusion

We evaluated three terrestrial ET models of different complexity in the ILAMB benchmarking system in China. Our results indicate that more complex models outperform simple models on the whole, as complex models marked highest ILAMB scores, though some exceptions exist. In terms of temporal simulation performance, the SWH model performed best for inter-annual variability simulation and ELM performed best for seasonal cycle simulation. For some special geographical environment regions, such as the Qinghai-Tibet Plateau and northwest region, models need to improve their ability to capture inter-annual variability and the seasonal cycle of ET. From the point of view of spatial distribution simulation, ELM and the SWH model are more closely related to the reference datasets, while the PT-JPL model performed poorly for the spatial distribution simulation of ET. In different PFT, the more complex models (ELM, SWH) performed better in forest, shrub and crop ecosystems and the simpler model (PT-JPL) performed better in grass ecosystems. We suggest that the performance difference may be due to different parameterizations and the simulation of important physical processes such as canopy resistance. This study provided a thorough evaluation of terrestrial ET models of different complexity by leveraging the strength of the ILAMB system. The approach will help guide efforts to understand the influence of model complexity on model performance and provide guidance on future directions of improving terrestrial ET models.

Acknowledgments

Support for this research was provided by National Key R&D Program of China

(2016YFC0501603), the National Natural Science Foundation of China (NSFC31961143022)， the National Natural Science Foundation of China (NSFC31922053). XC and TFK were supported by the Director, Office of Science, Office of Biological and Environmental Research of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 as part of their Regional and Global Climate Modeling program through the Reducing Uncertainties in Biogeochemical Interactions through Synthesis and Computation Scientific Focus Area (RUBISCO SFA) project. JBF contributed to this work from the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. California Institute of Technology. Government sponsorship acknowledged. Funding was provided in part by NASA programs: SUSMAP and ECOSTRESS. Copyright 2020. All rights reserved. XL, TFK and JBF TFK acknowledge support from the NASA Terrestrial Ecology Program IDS Award NNH17AE86I. This work used eddy covariance data acquired and shared by the FLUXNET community, including these networks: AmeriFlux, AfriFlux, AsiaFlux, CarboAfrica, CarboEuropeIP, CarboItaly, CarboMont, ChinaFlux, Fluxnet-Canada, GreenGrass, ICOS, KoFlux, LBA, NECC, OzFlux-TERN, TCOS-Siberia, and USCCC. The ERA-Interim reanalysis data are provided by ECMWF and processed by LSCE. The FLUXNET eddy covariance data processing and harmonization was carried out by the European Fluxes Database Cluster, AmeriFlux Management Project, and Fluxdata project of FLUXNET, with the support of CDIAC and ICOS Ecosystem Thematic Center, and the OzFlux, ChinaFlux and AsiaFlux offices. The dataset of FLUXNET is available at https://fluxnet.fluxdata.org/ website.

References

Adnan, R.M. et al., 2020. Reference Evapotranspiration Modeling Using New Heuristic Methods. Entropy, 22(5): 547.

Alizamir, M., Kisi, O., Muhammad Adnan, R. and Kuriqi, A., 2020. Modelling reference evapotranspiration by combining neuro-fuzzy and evolutionary strategies. Acta Geophysica: 1-14.

Badgley, G., Fisher, J.B., Jiménez, C., Tu, K.P. and Vinukollu, R., 2015. On uncertainty in global terrestrial evapotranspiration estimates from choice of input forcing datasets. Journal of Hydrometeorology, 16(4): 1449-1455.

Ball, J.T., Woodrow, I.E. and Berry, J.A., 1987. A model predicting stomatal conductance and its contribution to the control of photosynthesis under different environmental conditions, Progress in photosynthesis research. Springer, pp. 221-224.

Bonan, G. et al., 2013. Technical description of version 4.5 of the Community Land Model (CLM)(No. NCAR/TN-503+STR). doi:10.5065/D6RR1W7M.

Bonan, G.B. and Doney, S.C., 2018. Climate, ecosystems, and planetary futures: The challenge to predict life in Earth system models. Science, 359(6375).

Cai, X. et al., 2019. Improving representation of deforestation effects on evapotranspiration in the E3SM land model. Journal of Advances in Modeling Earth Systems, 11(8): 2412-2427.

Chaney, N.W., Herman, J.D., Ek, M.B. and Wood, E.F., 2016. Deriving global parameter estimates for the Noah land surface model using FLUXNET and machine learning. Journal of Geophysical Research: Atmospheres, 121(22): 13,218-13,235.

Chen, S. et al., 2009. Energy balance and partition in Inner Mongolia steppe ecosystems with different land use types. Agricultural and Forest Meteorology, 149(11): 1800-1809.

Chen, Y., Li, Z., Fan, Y., Wang, H. and Deng, H., 2015. Progress and prospects of climate change impacts on hydrology in the arid region of northwest China. Environmental research, 139: 11-9.

Collatz, G.J., Ribas-Carbo, M. and Berry, J., 1992. Coupled photosynthesis-stomatal conductance model for leaves of C4 plants. Functional Plant Biology, 19(5): 519-538.

Collier, N. et al., 2018. The International Land Model Benchmarking (ILAMB) System: Design, Theory, and Implementation. Journal of Advances in Modeling Earth Systems, 10(11): 2731-2754.

De Kauwe, M.G. et al., 2015. A test of an optimal stomatal conductance scheme within the CABLE land surface model. Geoscientific Model Development, 8(2): 431-452.

Ershadi, A., McCabe, M.F., Evans, J.P., Chaney, N.W. and Wood, E.F., 2014. Multi-site evaluation of terrestrial evaporation models using FLUXNET data. Agricultural and Forest Meteorology, 187: 46-61.

Eyring, V. et al., 2016. ESMValTool (v1. 0)–a community diagnostic and performance metrics tool for routine evaluation of Earth system models in CMIP. Geoscientific Model Development, 9: 1747-1802.

Farquhar, G.D., von Caemmerer, S.v. and Berry, J.A., 1980. A biochemical model of photosynthetic CO 2 assimilation in leaves of C 3 species. Planta, 149(1): 78-90.

Fisher, J.B. et al., 2017. The future of evapotranspiration: Global requirements for ecosystem functioning, carbon and climate feedbacks, agricultural management, and water resources. Water Resources Research, 53(4): 2618-2626.

Fisher, J.B., Tu, K.P. and Baldocchi, D.D., 2008. Global estimates of the land–atmosphere water flux based on monthly AVHRR and ISLSCP-II data, validated at 16 FLUXNET sites. Remote Sensing of Environment, 112(3): 901-919.

Fisher, J.B., Whittaker, R.J. and Malhi, Y., 2011. ET come home: potential evapotranspiration in geographical ecology. Global Ecology and Biogeography, 20(1): 1-18.

Gleckler, P. et al., 2016. A more powerful reality test for climate models. Eos, 97.

Guan, D.-X. et al., 2006. CO2 fluxes over an old, temperate mixed forest in northeastern China. Agricultural and Forest Meteorology, 137(3-4): 138-149.

Haughton, N. et al., 2016. The Plumbing of Land Surface Models: Is Poor Performance a Result of Methodology or Data Quality? Journal of Hydrometeorology, 17(6): 1705-1723.

Hobeichi, S., Abramowitz, G., Evans, J. and Ukkola, A., 2018. Derived Optimal Linear Combination Evapotranspiration (DOLCE): a global gridded synthesis ET estimate. Hydrology and Earth System Sciences (Online), 22(2).

Hogue, T.S., Bastidas, L.A., Gupta, H.V. and Sorooshian, S., 2006. Evaluating model performance and parameter behavior for varying levels of land surface model complexity. Water Resources Research, 42(8).

Hu, Z. et al., 2013. Modeling evapotranspiration by combing a two-source model, a leaf stomatal model, and a light-use efficiency model. Journal of Hydrology, 501: 186-192.

Hu, Z. et al., 2017. Modeling and Partitioning of Regional Evapotranspiration Using a Satellite-Driven Water-Carbon Coupling Model. Remote Sensing, 9(1): 54.

Jackson, R.D., 1985. Evaluating evapotranspiration at local and regional scales. Proceedings of the IEEE, 73(6): 1086-1096.

Jiménez, C. et al., 2011. Global inter-comparison of 12 land surface heat flux estimates. Journal of Geophysical Research, 116(D02102): doi:10.1029/2010JD014545.

Jung, M. et al., 2019. The FLUXCOM ensemble of global land-atmosphere energy fluxes. Scientific data, 6(1): 74.

Jung, M. et al., 2010. Recent decline in the global land evapotranspiration trend due to limited moisture supply. Nature, 467(7318): 951.

Kato, T. et al., 2006. Temperature and biomass influences on interannual changes in CO2 exchange in an alpine meadow on the Qinghai-Tibetan Plateau. Global change biology, 12(7): 1285-1298.

Khosa, F.V. et al., 2019. Evaluation of modeled actual evapotranspiration estimates from a land surface, empirical and satellite-based models using in situ observations from a South African semi-arid savanna ecosystem. Agricultural and Forest Meteorology, 279: 107706.

Leplastrier, M., 2002. Exploring the relationship between complexity and performance in a land surface model using the multicriteria method. Journal of Geophysical Research, 107(D20).

Lu, X., Chen, M., Liu, Y., Miralles, D.G. and Wang, F., 2017. Enhanced water use efficiency in global terrestrial ecosystems under increasing aerosol loadings. Agricultural and Forest Meteorology, 237-238: 39-49.

Luo, Y.Q. et al., 2012. A framework for benchmarking land models. Biogeosciences, 9(10): 3857-3874.

Ma, N., Szilagyi, J., Zhang, Y. and Liu, W., 2019a. Complementary-Relationship-Based Modeling of Terrestrial Evapotranspiration Across China During 1982–2012: Validations and Spatiotemporal Analyses. Journal of Geophysical Research: Atmospheres, 124(8): 4326-4351.

Ma, Y.-J. et al., 2019b. Evapotranspiration and its dominant controls along an elevation gradient in the Qinghai Lake watershed, northeast Qinghai-Tibet Plateau. Journal of Hydrology, 575: 257-268.

Mao, J. et al., 2015. Disentangling climatic and anthropogenic controls on global terrestrial evapotranspiration trends. Environmental Research Letters, 10(9): 094008.

Martens, B. et al., 2018. Towards Estimating Land Evaporation at Field Scales Using GLEAM. Remote Sensing, 10(11): 1720.

McCabe, M.F. et al., 2016. The GEWEX LandFlux project: evaluation of model evaporation using tower-based

and globally gridded forcing data. Geoscientific Model Development, 9(1): 283-305.
Medici, C., Wade, A.J. and Francés, F., 2012. Does increased hydrochemical model complexity decrease robustness? Journal of Hydrology, 440-441: 1-13.
Miner, G.L., Bauerle, W.L. and Baldocchi, D.D., 2017. Estimating the sensitivity of stomatal conductance to photosynthesis: a review. Plant, cell & environment, 40(7): 1214-1238.
Mingyue, C., Junbang, W., Shaoqiang, W., Hao, Y. and Yingnian, L., 2019. Temporal and Spatial Distribution of Evapotranspiration and Its Influencing Factors on Qinghai-Tibet Plateau from 1982 to 2014. Journal of Resources and Ecology, 10(2): 213-224.
Monteith, J.L., 1965. Evaporation and environment, Symposia of the society for experimental biology. Cambridge University Press (CUP) Cambridge, pp. 205-234.
Mueller, B. et al., 2013. Benchmark products for land evapotranspiration: LandFlux-EVAL multi-data set synthesis.
Mueller, B. et al., 2011. Evaluation of global observations-based evapotranspiration datasets and IPCC AR4 simulations. Geophysical Research Letters, 38(L06402): doi:10.1029/2010GL046230.
Oleson, K. et al., 2013. Technical Description of version 4.5 of the Community Land Model (CLM) Coordinating. BOULDER, COLORADO: 80307-3000.
Orth, R., Staudinger, M., Seneviratne, S.I., Seibert, J. and Zappa, M., 2015. Does model performance improve with complexity? A case study with three hydrological models. Journal of Hydrology, 523: 147-159.
Pastorello, G. et al., 2017. A new data set to keep a sharper eye on land-air exchanges. Eos, Transactions American Geophysical Union (Online), 98(8).
Peaucelle, M. et al., 2019. Covariations between plant functional traits emerge from constraining parameterization of a terrestrial biosphere model. Global Ecology and Biogeography, 28(9): 1351-1365.
Polhamus, A., Fisher, J.B. and Tu, K.P., 2013. What controls the error structure in evapotranspiration models? Agricultural and forest meteorology, 169: 12-24.
Priestley, C.H.B. and Taylor, R., 1972. On the assessment of surface heat flux and evaporation using large-scale parameters. Monthly weather review, 100(2): 81-92.
Randerson, J.T. et al., 2009. Systematic assessment of terrestrial biogeochemistry in coupled climate–carbon models. Global change biology, 15(10): 2462-2484.
Running, S.W. et al., 2004. A continuous satellite-derived measure of global terrestrial primary production. Bioscience, 54(6): 547-560.
Schwalm, C.R. et al., 2013. Sensitivity of inferred climate model skill to evaluation decisions: a case study using CMIP5 evapotranspiration. Environmental Research Letters, 8(2): 024028.
Shan, N. et al., 2019. Modeling canopy conductance and transpiration from solar-induced chlorophyll fluorescence. Agricultural and Forest Meteorology, 268: 189-201.
Shi, P. et al., 2006. Net ecosystem CO2 exchange and controlling factors in a steppe—Kobresia meadow on the Tibetan Plateau. Science in China Series D: Earth Sciences, 49(S2): 207-218.
Shuttleworth, W.J. and Wallace, J., 1985. Evaporation from sparse crops-an energy combination theory. Quarterly Journal of the Royal Meteorological Society, 111(469): 839-855.
Stofferahn, E. et al., 2019. The Arctic-Boreal vulnerability experiment model benchmarking system. Environmental Research Letters, 14(5): 055002.
Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. Journal of Geophysical Research: Atmospheres, 106(D7): 7183-7192.
Vinukollu, R.K., Wood, E.F., Ferguson, C.R. and Fisher, J.B., 2011. Global estimates of evapotranspiration for

climate studies using multi-sensor remote sensing data: Evaluation of three process-based approaches. Remote Sensing of Environment, 115: 801-823.
Wang, X., Yang, M., Pang, G., Wan, G. and Chen, X., 2014. Simulation and improvement of land surface processes in Nameqie, Central Tibetan Plateau, using the Community Land Model (CLM3.5). Environmental Earth Sciences, 73(11): 7343-7357.
Wang, Y. et al., 2017. Changes in mean and extreme temperature and precipitation over the arid region of northwestern China: Observation and projection. Advances in Atmospheric Sciences, 34(3): 289-305.
Yang, P., Xia, J., Zhang, Y., Zhan, C. and Qiao, Y., 2018. Comprehensive assessment of drought risk in the arid region of Northwest China based on the global palmer drought severity index gridded data. Science of the Total Environment, 627: 951-962.
Yu, G.-R. et al., 2006. Overview of ChinaFLUX and evaluation of its eddy covariance measurement. Agricultural and Forest Meteorology, 137(3-4): 125-137.
Zhang, L., Luo, Y., Yu, G. and Zhang, L., 2010. Estimated carbon residence times in three forest ecosystems of eastern China: Applications of probabilistic inversion. Journal of Geophysical Research, 115(G1).
Zhang, Y. et al., 2016. Multi-decadal trends in global terrestrial evapotranspiration and its components. Scientific reports, 6: 19124.

Appendix A. List of abbreviations and acronyms

DBF	deciduous broadleaf forest
DNF	deciduous needleleaf forest
DOLCE	Derived Optimal Linear Combination Evapotranspiration
E3SM	Energy Exascale Earth System Model
EBF	evergreen broadleaf forest
ELM	Energy Exascale Earth System Model Land Model
ENF	evergreen needleleaf forest
ESMs	earth system models
ET	evapotranspiration
GLEAM	Global Land Evaporation Amsterdam Model
GSWP3	Global Soil Wetness Project Phase 3
IAV	inter-annual variability
ILAMB	International Land Model Benchmarking
MF	mixed forest
NDVI	Normalized Difference Vegetation Index
PFT	plant functional types
PM	plateau and mountain climate
PT-JPL	Priestley Taylor-Jet Propulsion Laboratory
r	correlation
RMSE	root mean square error
R_n	net radiation
SC	seasonal cycle

SD	standard deviation
SM	subtropical monsoon climate
SWH	Shuttleworth Wallace Hu
TC	temperate continental climate
TM	temperate monsoon climate