

Lawrence Berkeley National Laboratory

Recent Work

Title

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

Permalink

<https://escholarship.org/uc/item/55j7j6wq>

Authors

Wu, G
Cai, X
Keenan, TF
[et al.](#)

Publication Date

2020-11-01

DOI

10.1016/j.jhydrol.2020.125553

Peer reviewed

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

3
4 Genan Wu^{1,2,3,4}, Xitian Cai³, Trevor F. Keenan^{3,4}, Shenggong Li^{1,2,*}, Xiangzhong Luo^{3,4}, Joshua B. Fisher⁵,
5 Ruochen Cao⁶, Fa Li^{3,7}, Adam J Purdy⁵, Wei Zhao^{1,6}, Xiaomin Sun^{1,2}, Zhongmin Hu^{6,8,*}

6 1 Synthesis Research Center of Chinese Ecosystem Research Network, Key Laboratory of Ecosystem Network Observation
7 and Modeling, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing,
8 China

9 2 College of Resources and Environment, University of Chinese Academy of Sciences, Beijing, China

10 3 Lawrence Berkeley National Laboratory, Berkeley, CA, USA

11 4 UC Berkeley, Berkeley, CA, USA

12 5 Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA

13 6 School of Geography, South China Normal University, Guangzhou, China

14 7 State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, China

15 8 Southern Marine Science and Engineering Guangdong Laboratory, Zhuhai, China.

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

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

18
19 **Abstract:** Land surface models range in complexity of terrestrial evapotranspiration, yet it is
20 unknown how model complexity translates to accuracy of modeled evapotranspiration
21 estimates. Here, we use the International Land Model Benchmarking system to assess ET
22 estimates from three models of varying complexity driven by the same forcing datasets: an
23 earth system model, a terrestrial biosphere model, and a stand-alone ET model. The
24 performance assessment includes both temporal and spatial evaluation, and different plant
25 functional types across China. Our results indicate that the most complex model, an earth
26 system model, performed best against the benchmarking datasets and metrics. Terrestrial

27 biosphere model performed best in simulating inter-annual variability of ET, while earth
28 system model performed best in simulating the seasonal cycle. The more complex models
29 (earth system model and terrestrial biosphere model) perform better in forest, shrub and crop
30 ecosystems, while the simpler model (stand-alone ET model) perform better in grass
31 ecosystems. Our study demonstrates the impact of model complexity on ET estimates and
32 highlights directions for future ET model improvements.

33 **Key words:** Benchmarking, evapotranspiration model, model complexity

34 **1. Introduction**

35 Evapotranspiration (ET) is a key component of the global water budget and is crucial to
36 agriculture and water management, the sustainability of ecosystems, and the water and carbon
37 exchanges between land and atmosphere (Fisher et al., 2017). However, the estimation of
38 large-scale ET from ground-based measurements alone remains challenging due to the sparse
39 network of point observations and the high spatial and temporal variability of ET (Lu et al.,
40 2017). To address this limitation, various terrestrial ET models have been developed (Jiménez
41 et al., 2011; McCabe et al., 2016; Mueller et al., 2011; Vinukollu et al., 2011).

42

43 Terrestrial ET models play a vital role in diagnosing and predicting global water fluxes and in
44 evaluating the impacts of changing climate (Mao et al., 2015). In recent years, a variety of
45 physical process models have been developed to estimate the spatial distribution of
46 evapotranspiration (ET) at various scales ranging from the stand scale to global. From
47 empirical and semi-empirical method (i.e. Jackson model, Priestley-Taylor model) to
48 physical processed method (i.e. Shuttleworth-Wallace model, Community Land Model),

49 much progress has been made incorporate more physical processes into ET simulations
50 (Bonan et al., 2013; Jackson, 1985; Priestley and Taylor, 1972; Shuttleworth and Wallace,
51 1985). In addition, some statistic and machine learning methods were used to improve ET
52 models performance and accuracy (Adnan et al., 2020; Alizamir et al., 2020). As ET models
53 become increasingly complex and the number of model parameters rapidly expands, there is a
54 growing need for a comprehensive and multifaceted evaluation of the performance of models
55 of different levels of complexity (Haughton et al., 2016; Hogue et al., 2006). In this study,
56 “complexity” is defined in terms of the number of process-related variables and parameters
57 and the hierarchy of model structure. In terrestrial ET models, for example, the
58 Priestley-Taylor model (Priestley and Taylor, 1972)—a simplification of the
59 Penman-Monteith equation (Monteith, 1965)—requires less forcing data and thus does not
60 consider explicitly the impact of vapor pressure deficit (VPD) or canopy resistance. This
61 method is convenient to use in the absence of detailed meteorological measurements. By
62 contrast, the Penman-Monteith model and the Shuttleworth-Wallace model (Shuttleworth and
63 Wallace, 1985) consider complex biogeochemical and biogeophysical land surface processes
64 and therefore require more meteorological measurements and parameters (Fisher et al., 2011).
65 Specifically, the Shuttleworth-Wallace model partitions ET into soil water evaporation and
66 plant transpiration and contains more complexity estimation of ET processes.

67

68 In recent decades, earth system models (ESM) which simulate biogeochemical processes on
69 the land surface, which are fully coupled with physical climate simulations, have been
70 developed rapidly and widely used (Bonan and Doney, 2018). Meanwhile, the estimation of

71 the physical-process variables of an ESM such as ET is becoming increasingly
72 comprehensive and sophisticated. Compared to other terrestrial ET models, ESM require
73 higher temporal-spatial resolution forcing data and physical parameters (Mueller et al., 2013).
74 Although more complicated ET models can provide more details involved in
75 atmosphere-terrestrial water exchange, they are also potentially prone to greater uncertainties
76 propagated from other related processes (Orth et al., 2015). There remains a lack of
77 knowledge on the optimal complexity of ET models on the regional scale.

78

79 Model benchmarking has emerged as an effective approach to evaluate model performance
80 relative to multiple observational constraints as well as other models (Collier et al. 2018).
81 Most recently, the International Land Model Benchmarking (ILAMB) System (Collier et al.,
82 2018; Luo et al., 2012; Stofferahn et al., 2019), the ESM Evaluation Tool (Eyring et al., 2016),
83 the Program for Climate Model Diagnosis and Intercomparison Metrics Package (Gleckler et
84 al., 2016) and other benchmarking system were created to explore land surface model
85 intercomparison and facilitate internationally accepted benchmarks (Schwalm et al., 2013).

86

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

93 for each of the three ET models, in order to limit the uncertainty of the forcing data (Badgley
94 et al., 2015) and focus on the effect of model complexity.

95

96 2. Methodology

97 2.1 ILAMB Description

98 As land surface models become increasingly complex and observational data volumes rapidly
99 expand, there is a growing need for comprehensive and multifaceted evaluation of model
100 fidelity. Building on past model evaluation work (Randerson et al., 2009), Luo et al. (2012)
101 and Collier et al. (2018) developed an extensible model benchmarking package in support of
102 the goals of the International Land Model Benchmarking (ILAMB) activity. The ILAMB
103 benchmarking system compares model estimates against the best-available observations and
104 observation-based extrapolations, including atmosphere CO₂ concentrations, surface fluxes,
105 hydrology, soil carbon and nutrient biogeochemistry, ecosystem processes and states, and
106 vegetation dynamics.

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

$$111 \quad \text{bias}(\mathbf{x}) = \overline{v_{\text{mod}}}(\mathbf{x}) - \overline{v_{\text{ref}}}(\mathbf{x}) \quad (1)$$

112 The variable \mathbf{x} is spatial domain which represents the areas created by cell boundaries or the
113 areas connected with data sites. $\overline{v_{\text{mod}}}(\mathbf{x})$ is the mean value over time of a modelled dataset.
114 $\overline{v_{\text{ref}}}(\mathbf{x})$ is the mean value over time of a reference dataset. We then nondimensionalized the
115 biases into a relative error using the centralized RMS (Root Mean Square) of the reference
116 dataset following equation (2):

$$117 \quad \text{crms}(x) = \sqrt{\frac{1}{t_f - t_0} \int_{t_0}^{t_f} (v_{\text{ref}}(t, x) - \overline{v_{\text{ref}}}(x))^2 dt} \quad (2)$$

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

$$120 \quad \varepsilon_{\text{bias}}(x) = |\text{bias}(x)|/\text{crms}(x) \quad (3)$$

121 The bias score as a function of space is:

$$122 \quad S_{\text{bias}}(x) = e^{-\varepsilon_{\text{bias}}(x)} \quad (4)$$

123 And the scalar score

$$124 \quad S_{\text{bias}} = \overline{S_{\text{bias}}(x)} \quad (5)$$

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

$$127 \quad \text{RMSE}(x) = \sqrt{\frac{1}{t_f - t_0} \int_{t_0}^{t_f} (v_{\text{mod}}(t, x) - v_{\text{ref}}(t, x))^2 dt} \quad (6)$$

128 To score the RMSE, we use the methods similar to Eq. (2-5). Please refer to Collier et al.
 129 (2018) for more details. ILAMB evaluates the phase shift of the annual cycle of data sets that
 130 have intra-annual variability by comparing the timing of the maximum value in a year, $c(v)$
 131 within each. Then, we approximate the phase shift from the reference to model data sets by
 132 subtracting their respective $c(v)$,

$$133 \quad \theta(x) = \arg \max_t (c_{\text{mod}}(t, x)) - \arg \max_t (c_{\text{ref}}(t, x)) \quad (7)$$

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

$$136 \quad S_{\text{phase}}(x) = \frac{1}{2} (1 + \cos(\frac{2\pi}{365} \theta(x))) \quad (8)$$

137 And the scalar score is:

$$138 \quad S_{\text{phase}} = \overline{S_{\text{phase}}(x)} \quad (9)$$

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

$$141 \quad iav_{\text{ref}}(x) = \sqrt{\frac{1}{t_f - t_0} \int_{t_0}^{t_f} (v_{\text{ref}}(t, x) - c_{\text{ref}}(t, x))^2 dt} \quad (10)$$

$$142 \quad iav_{\text{mod}}(x) = \sqrt{\frac{1}{t_f - t_0} \int_{t_0}^{t_f} (v_{\text{mod}}(t, x) - c_{\text{mod}}(t, x))^2 dt} \quad (11)$$

$$143 \quad \varepsilon_{\text{iav}}(x) = (iav_{\text{mod}}(x) - iav_{\text{ref}}(x)) / iav_{\text{ref}}(x) \quad (12)$$

144 and then computing a score as a function of space,

$$145 \quad s_{\text{iav}}(x) = e^{-\varepsilon_{\text{iav}}(x)} \quad (13)$$

146 The scalar score is estimated by:

$$147 \quad S_{\text{iav}} = \overline{s_{\text{iav}}(x)} \quad (14)$$

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

$$150 \quad \sigma = \frac{\text{stdev}(\overline{v_{\text{mod}}(x)})}{\text{stdev}(\overline{v_{\text{ref}}(x)})} \quad (15)$$

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

$$153 \quad S_{\text{dist}} = \frac{2(1+R)}{(\sigma + \frac{1}{\sigma})^2} \quad (16)$$

154 Where the main idea is that we penalize the score when R and σ deviate from a value of 1.

155 The overall score for a given variable and data product is a composite of the suite of metrics
 156 defined above. We use a weighted sum,

$$157 \quad S_{\text{overall}} = \frac{S_{\text{bias}} + 2S_{\text{rmse}} + S_{\text{phase}} + S_{\text{iav}} + S_{\text{dist}}}{1+2+1+1+1} \quad (17)$$

158 Where the RMSE score is doubled to emphasize its importance. In addition, we show the
 159 relative score (i.e., Z score), indicating which models or model versions perform better with

160 respect to others contained in the overall analysis. More details of the underlying metrics are
161 available in Collier et al. (2018).

162 **2.2 Data Sets**

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

174

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

182 the ET variable is estimated as:

$$183 \quad w_{\text{FLUXNET}}^{\text{ET}} = \frac{3 \times 5}{3 \times 5 + 3 \times 5 + 3 \times 5 + 3 \times 5} \approx 25\% \quad (18)$$

184

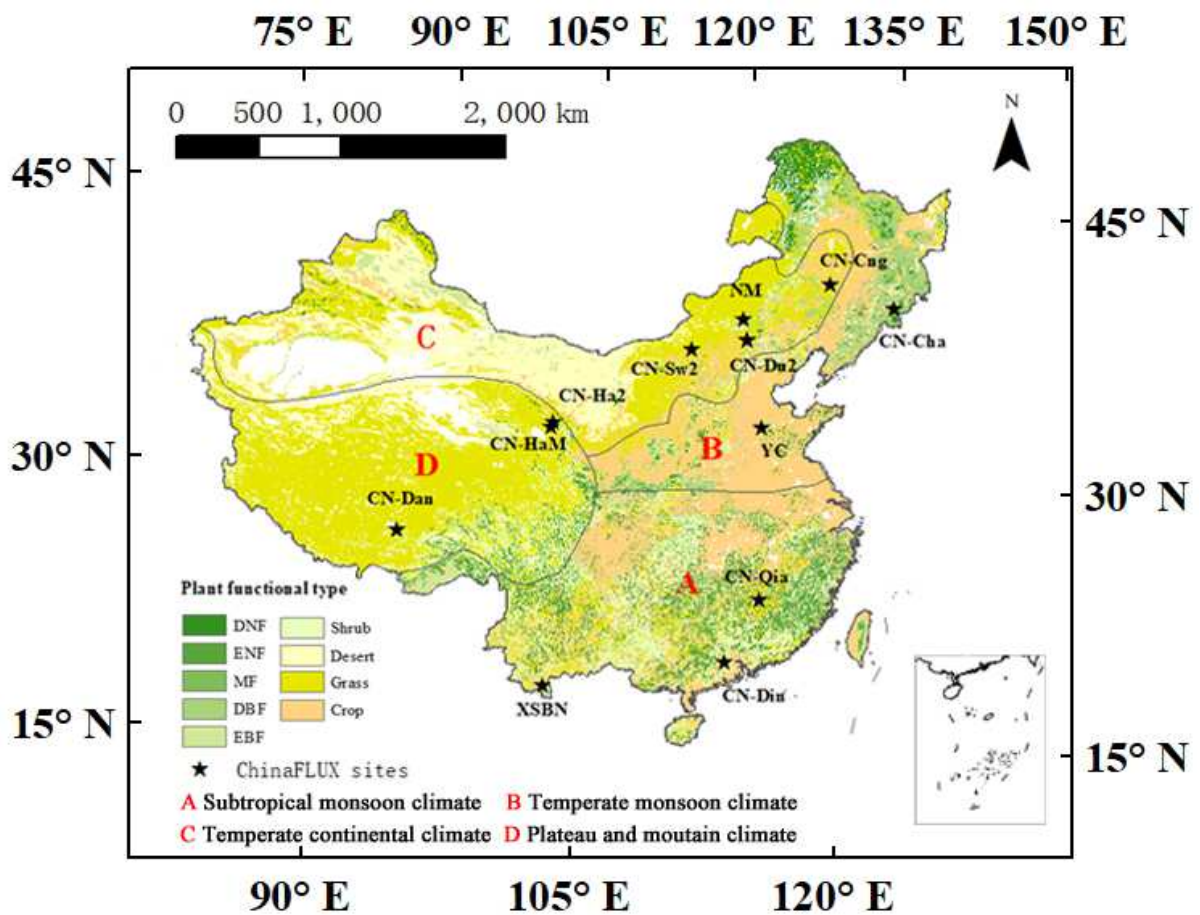
185 **Table 1.** References and weighting of evapotranspiration (ET) data sets used to blend the
186 overall score.

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)

187

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

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



205
 206 **Figure 1.** Locations of the 12 ChinaFLUX sites and distribution of plant functional type and
 207 climate zones.

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

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

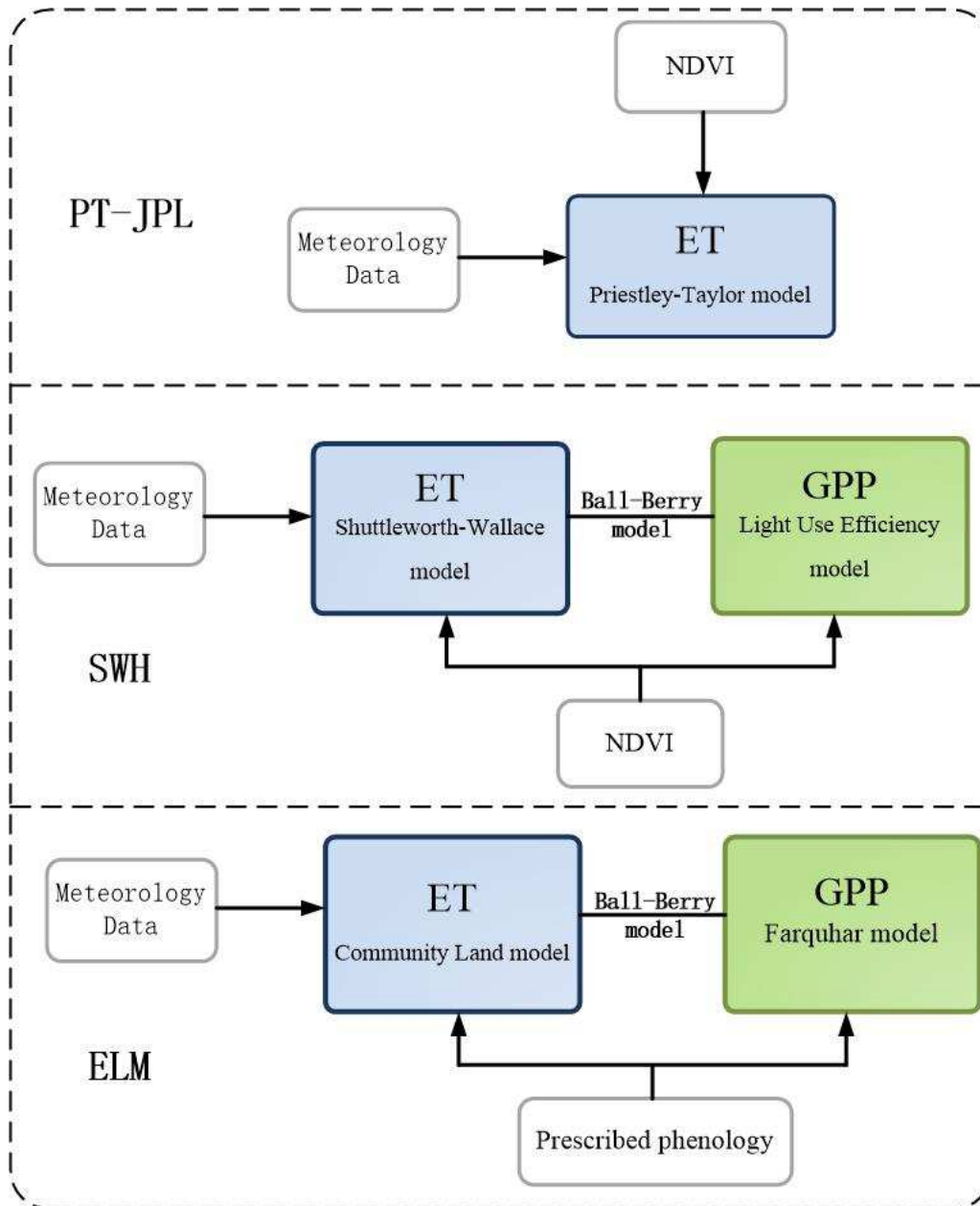
CN-Cha	MF	42.4	128.1	2005-2014	Guan et al. (2006)
CN-Cng	GRA	44.59	123.51	2007-2010	-
CN-Dan	GRA	30.50	91.07	2004-2008	Shi et al. (2006)
CN-Din	EBF	23.17	112.54	2003-2005	Zhang et al. (2010)
CN-Du2	GRA	42.05	116.28	2006-2008	Chen et al. (2009)
CN-Ha2	WET	37.61	101.33	2003-2005	-
CN-HaM	GRA	37.37	101.18	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)

209

210 **2.3 ET Model Descriptions**

211 To limit the uncertainty of the forcing data and focus on the effect of different model
212 complexity, we used the same meteorology datasets from 1980 to 2010 (GSWP3,
213 <https://www.isimip.org/gettingstarted/details/4/>) and satellite remote sensing datasets
214 (Normalized Difference Vegetation Index (NDVI) GIMMS product,
215 <https://glam1.gsfc.nasa.gov/>) to run the three models. The simplest ET model is the Priestley
216 Taylor-Jet Propulsion Laboratory (PT-JPL) model which is developed from Priestley-Taylor
217 model (Fisher et al., 2008; Priestley and Taylor, 1972). The PT-JPL model incorporates a
218 variety of data sources from meteorological data (i.e., net radiation (R_n), air temperature,

219 vapor pressure) and satellite observations (NDVI, visible spectrum reflectance, near-infrared
220 spectrum reflectance). We use the Shuttleworth-Wallace-Hu (SWH) model as a representative
221 of intermediate complex models (Hu et al., 2013; Hu et al., 2017), which is developed based
222 on the Shuttleworth-Wallace model and coupled light use efficiency model (Shuttleworth and
223 Wallace, 1985). Meteorological data (i.e., air temperature, precipitation, relative humidity,
224 wind speed, and R_n) and satellite products (i.e., NDVI) are the forcing data for the SWH
225 model. We used the version 1 of the Energy Exascale Earth System Model (E3SM) Land
226 Model (ELMv1) as a representative of the most complex ET model, which was branched
227 from the version 4.5 of the Community Land Model (CLM4.5; Oleson et al. (2013)) with a
228 specific version tag 4_5_71 (Cai et al., 2019). The forcing fields include surface air
229 temperature, precipitation, wind speed, relative humidity, surface pressure, incoming solar
230 radiation, and incoming longwave radiation. (Figure 2)



231

232 **Figure 2.** Evapotranspiration models: Priestley Taylor-Jet Propulsion Laboratory (PT-JPL)

233 model, Shuttleworth-Wallace-Hu (SWH) model, and Energy Exascale Earth System Model

234 Land Model (ELM).

235

236 **3. Results**

237 **3.1 Overall performance**

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

244

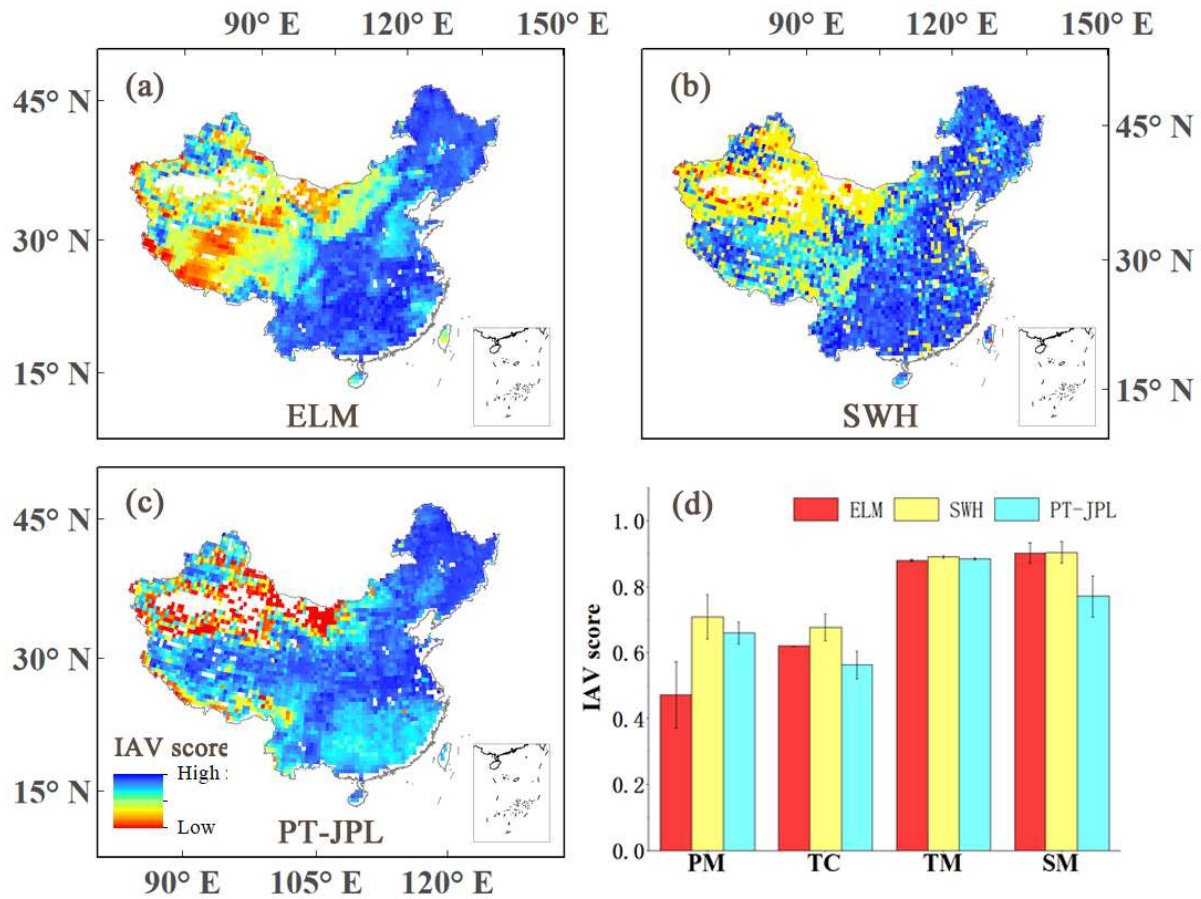
245 **3.2 Inter-annual variability and seasonal cycle simulation performance**

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

252

253 For the different climate region in China (Figure 3d), ELM model had the lowest score in
254 simulating the inter-annual variability of ET in the plateau and mountain climate region
255 (mean score = 0.47). There is a need to improve the ET inter-annual variability simulation of
256 the three terrestrial ET models in the temperate continental climate region (mean score: ELM
257 = 0.62, SWH = 0.68, JPL = 0.56). All three ET models perform equally well in the temperate

258 monsoon climate region (mean score: ELM = 0.88, SWH = 0.89, JPL = 0.88). In the
 259 subtropical monsoon climate region, PT-JPL model had the worst performance of ET
 260 inter-annual variability simulation (mean score = 0.77).
 261



262
 263 **Figure 3.** The spatial distribution of inter-annual variability (IAV) score of three models: (a)
 264 ELM, (b) SWH and (c) PT-JPL and (d) the inter-annual variability score in different climate
 265 change: plateau and mountain climate (PM), temperate continental climate (TC), temperate
 266 monsoon climate (TM), subtropical monsoon climate (SM).

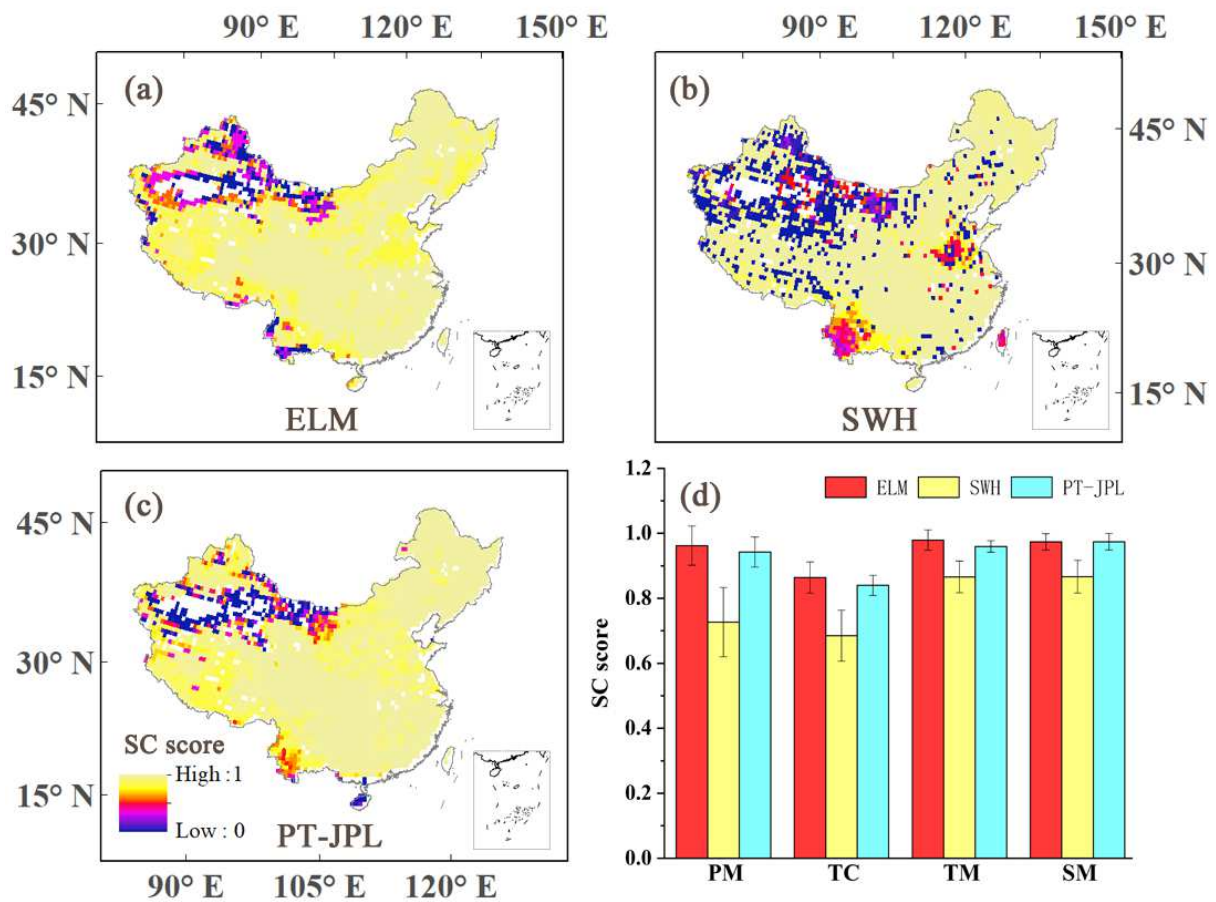
267
 268 In terms of seasonal cycle score, which compares the timing of the maximum ET of the
 269 annual cycle between reference dataset and model dataset, ELM and PT-JPL (mean

270 score=0.91, 0.90) performs better than SWH model (mean score=0.78). In northwestern and
 271 southwestern of China, the simulation of seasonal cycle of the three ET models had lower
 272 scores especially the SWH model (Figure 4).

273

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

279



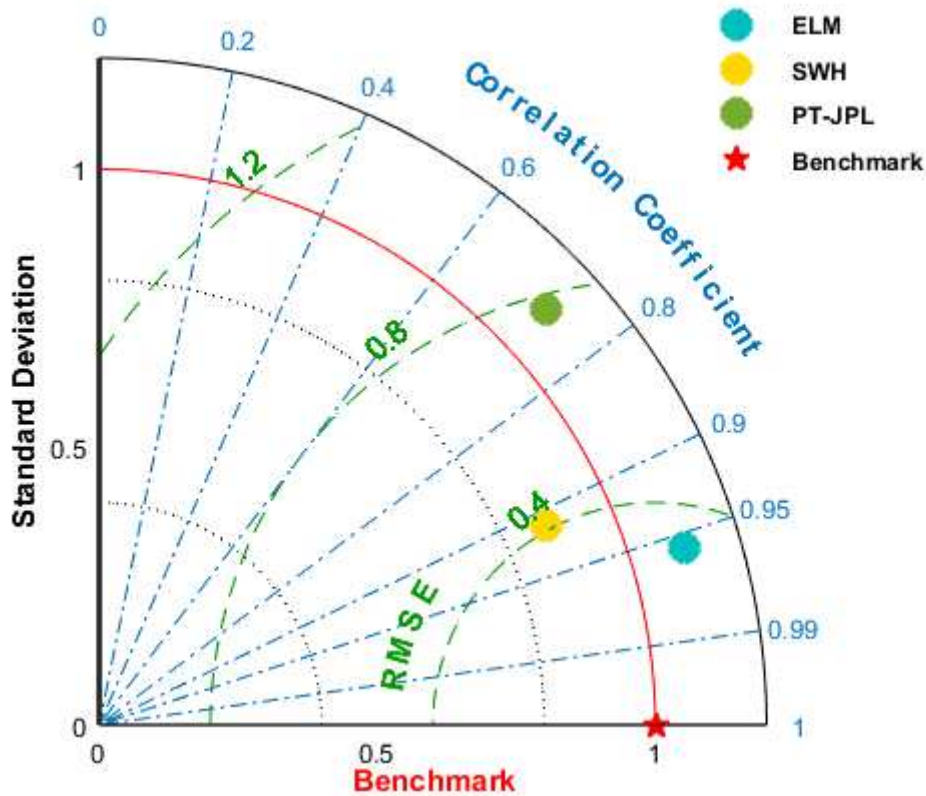
280

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

283

284 **3.3 Spatial variability performance**

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



302

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

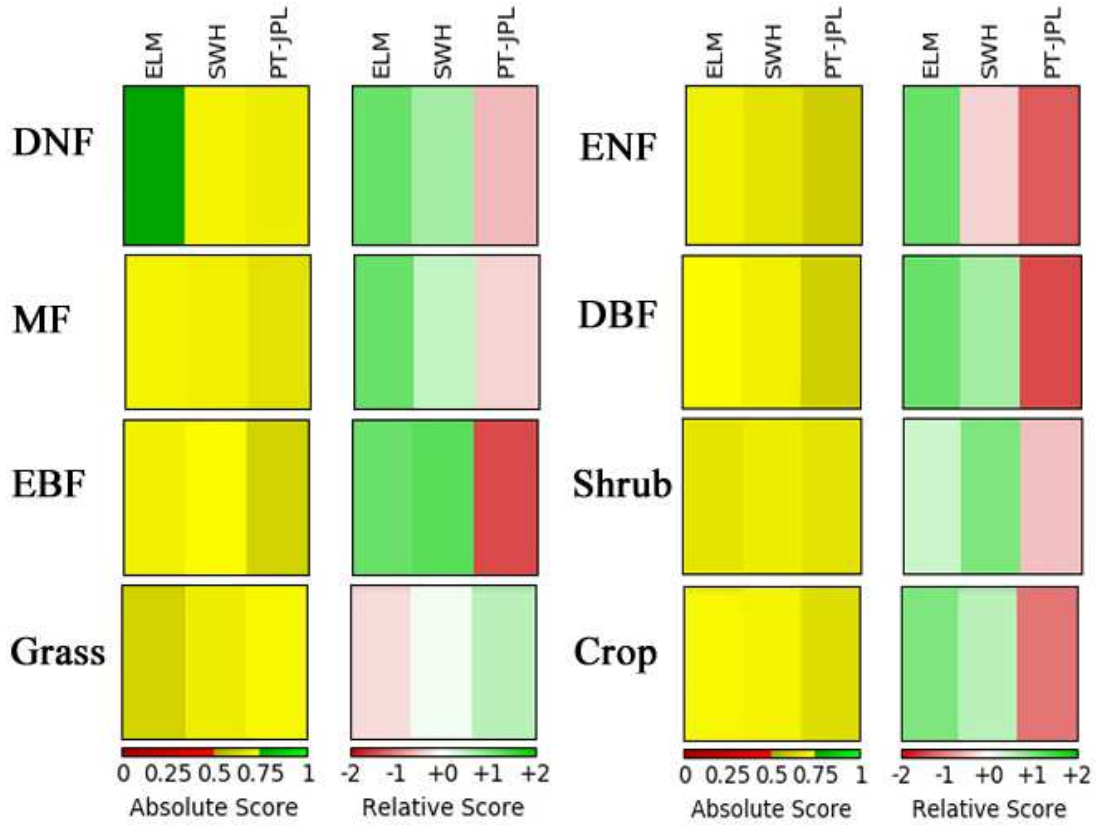
305

306 **3.4 Model performance in different plant functional types**

307 In different plant functional types (PFT), the three levels of complexity terrestrial ET models
 308 have different performance relative to the reference datasets. The most complicated ET model,
 309 ELM, shows the best performance in DNF, ENF, MF, DBF, and Crop (overall score = 0.75,
 310 0.69, 0.70, 0.72, 0.71) but performs worst in Grass (overall score = 0.61). The best
 311 performance of the intermediate complexity model, SWH is achieved in EBF and Shrub
 312 (overall score = 0.72, 0.69). And the simplest model, PT-JPL have the best performance in
 313 Grass (overall score = 0.71). Both of SWH and PT-JPL models has poor performance in forest

314 ecosystems. Additionally, the relative score revealed that PT-JPL model perform worse in
 315 ENF, DBF, and EBF compared to the other models. (Figure 6)

316



317

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

320

321 **4. Discussion**

322 **4.1 Overall performance of the three levels of complexity terrestrial ET models**

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

339

340 **4.2 Temporal and spatial simulation performance**

341 Given that direct model evaluation is possible only with contemporary *in-situ* observations, it
342 is difficult to assess the models' capacities to capture spatial variation at large scale. Khosa et

343 al. (2019) evaluated and calibrated surface, empirical and satellite-based models performance
344 including inter-annual variation and seasonal cycle performance compared with in situ ET
345 measurement in South Africa. Ma et al. (2019a) validated a 31-year ET product by using
346 plot-scale eddy covariance measurement and basin-scale water-balance-derived
347 evapotranspiration rates and quantified the spatial and temporal variability of ET in China.
348 However, we still lack a quantitative assessment of ET model performance distribution for
349 inter annual variability and seasonal cycle. In this study, we leveraged the ILAMB system to
350 enable improved testing of multiple terrestrial ET models, which used a wide variety of
351 regional-scale gridded observations, site specific observations, and integrative observations to
352 allow a more robust model benchmarking framework.

353

354 As shown in Figure 3, SWH performs best in terms of inter-annual variability simulation.
355 And the simulation of inter-annual variability of the three ET models (ELM, SWH, PT-JPL)
356 is poor in the northwest of China (temperate continental climate region). In the northwest arid
357 region, temperature and precipitation experienced a sharp increasing in the past 50 years
358 (Yang et al., 2018). The precipitation trend changed in 1987, and since then has been in a
359 state of high volatility. Temperature experienced a “sharp” increase in 1997; since then, it has
360 remained highly volatile, and the increasing trend slowed (Chen et al., 2015; Wang et al.,
361 2017). Meanwhile, whether reanalysis climate product or interpolation climate data is
362 effected by in situ measurements which is less distributed in the northwest of China. These
363 may be one of the reasons for the poor inter-annual variability simulation performance in the
364 northwest of China.

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

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

382

383 In term of the spatial distribution simulation, ELM and SWH models have higher correlation
384 coefficients with the reference dataset (0.96, 0.91, respectively), which is higher than the
385 coefficient for PT-JPL model (0.72). On the other hand, ELM and the SWH model showed
386 the smaller RMSE in comparison with the benchmark data. Considering the evidence above,

387 we found that the more complex models (ELM, SWH) perform better for the ET spatial
388 distribution than the simpler model (PT-JPL). A possible explanation for these results may be
389 some key parameters of terrestrial ET model are space-time scale dependent and relate to
390 traits in specific environmental (Chaney et al., 2016; Peaucelle et al., 2019). For the more
391 complex models (ELM, SWH), the variations of key parameters are considered in the
392 physical-process simulation in different PFT. It is therefore likely that the more complex
393 models simulate spatial distribution better in China, due to their ability to better consider the
394 variations and diversity in the ecosystem characteristics.

395

396 **4.3 Model performance in different plant functional types**

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

401

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

409 (Ball et al., 1987) to calculate stomatal resistance. SWH used a light use efficiency model
410 (Running et al., 2004) to estimate the photosynthesis rate, which is a key parameter in the
411 Ball-Berry model, while the photosynthesis rate in ELM is based on biochemical models
412 (Collatz et al., 1992; Farquhar et al., 1980). ET integrates biochemical and biophysical land
413 surface processes between the Earth's surface and atmosphere (Jung et al., 2010; Zhang et al.,
414 2016). Coupling biochemical and biophysical processes in terrestrial ET models is thus
415 expected to lead to improved performance. This improved process representation could
416 explain why the ELM model performs better in particular in forest ecosystems, which have a
417 more complex canopy structure.

418

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

424

425 **5. Conclusion**

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

444

445 ***Acknowledgments***

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

447 (2016YFC0501603), the National Natural Science Foundation of China (NSFC31961143022),
448 the National Natural Science Foundation of China (NSFC31922053). XC and TFK were
449 supported by the Director, Office of Science, Office of Biological and Environmental
450 Research of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 as
451 part of their Regional and Global Climate Modeling program through the Reducing
452 Uncertainties in Biogeochemical Interactions through Synthesis and Computation Scientific
453 Focus Area (RUBISCO SFA) project. JBF contributed to this work from the Jet Propulsion
454 Laboratory, California Institute of Technology, under a contract with the National Aeronautics
455 and Space Administration. California Institute of Technology. Government sponsorship
456 acknowledged. Funding was provided in part by NASA programs: SUSMAP and
457 ECOSTRESS. Copyright 2020. All rights reserved. XL, TFK and JBF TFK acknowledge
458 support from the NASA Terrestrial Ecology Program IDS Award NNH17AE86I. This work
459 used eddy covariance data acquired and shared by the FLUXNET community, including these
460 networks: AmeriFlux, AfriFlux, AsiaFlux, CarboAfrica, CarboEuropeIP, CarboItaly,
461 CarboMont, ChinaFlux, Fluxnet-Canada, GreenGrass, ICOS, KoFlux, LBA, NECC,
462 OzFlux-TERN, TCOS-Siberia, and USCCC. The ERA-Interim reanalysis data are provided
463 by ECMWF and processed by LSCE. The FLUXNET eddy covariance data processing and
464 harmonization was carried out by the European Fluxes Database Cluster, AmeriFlux
465 Management Project, and Fluxdata project of FLUXNET, with the support of CDIAC and
466 ICOS Ecosystem Thematic Center, and the OzFlux, ChinaFlux and AsiaFlux offices. The
467 dataset of FLUXNET is available at <https://fluxnet.fluxdata.org/> website.

468

469 **References**

- 470 Adnan, R.M. et al., 2020. Reference Evapotranspiration Modeling Using New Heuristic Methods. *Entropy*,
471 22(5): 547.
- 472 Alizamir, M., Kisi, O., Muhammad Adnan, R. and Kuriqi, A., 2020. Modelling reference evapotranspiration by
473 combining neuro-fuzzy and evolutionary strategies. *Acta Geophysica*: 1-14.
- 474 Badgley, G., Fisher, J.B., Jiménez, C., Tu, K.P. and Vinukollu, R., 2015. On uncertainty in global terrestrial
475 evapotranspiration estimates from choice of input forcing datasets. *Journal of Hydrometeorology*, 16(4):
476 1449-1455.
- 477 Ball, J.T., Woodrow, I.E. and Berry, J.A., 1987. A model predicting stomatal conductance and its contribution to
478 the control of photosynthesis under different environmental conditions, *Progress in photosynthesis*
479 *research*. Springer, pp. 221-224.
- 480 Bonan, G. et al., 2013. Technical description of version 4.5 of the Community Land Model (CLM)(No.
481 NCAR/TN-503+STR). doi:10.5065/D6RR1W7M.
- 482 Bonan, G.B. and Doney, S.C., 2018. Climate, ecosystems, and planetary futures: The challenge to predict life in
483 Earth system models. *Science*, 359(6375).
- 484 Cai, X. et al., 2019. Improving representation of deforestation effects on evapotranspiration in the E3SM land
485 model. *Journal of Advances in Modeling Earth Systems*, 11(8): 2412-2427.
- 486 Chaney, N.W., Herman, J.D., Ek, M.B. and Wood, E.F., 2016. Deriving global parameter estimates for the Noah
487 land surface model using FLUXNET and machine learning. *Journal of Geophysical Research:*
488 *Atmospheres*, 121(22): 13,218-13,235.
- 489 Chen, S. et al., 2009. Energy balance and partition in Inner Mongolia steppe ecosystems with different land use
490 types. *Agricultural and Forest Meteorology*, 149(11): 1800-1809.
- 491 Chen, Y., Li, Z., Fan, Y., Wang, H. and Deng, H., 2015. Progress and prospects of climate change impacts on
492 hydrology in the arid region of northwest China. *Environmental research*, 139: 11-9.
- 493 Collatz, G.J., Ribas-Carbo, M. and Berry, J., 1992. Coupled photosynthesis-stomatal conductance model for
494 leaves of C4 plants. *Functional Plant Biology*, 19(5): 519-538.
- 495 Collier, N. et al., 2018. The International Land Model Benchmarking (ILAMB) System: Design, Theory, and
496 Implementation. *Journal of Advances in Modeling Earth Systems*, 10(11): 2731-2754.
- 497 De Kauwe, M.G. et al., 2015. A test of an optimal stomatal conductance scheme within the CABLE land surface
498 model. *Geoscientific Model Development*, 8(2): 431-452.
- 499 Ershadi, A., McCabe, M.F., Evans, J.P., Chaney, N.W. and Wood, E.F., 2014. Multi-site evaluation of terrestrial
500 evaporation models using FLUXNET data. *Agricultural and Forest Meteorology*, 187: 46-61.
- 501 Eyring, V. et al., 2016. ESMValTool (v1. 0)—a community diagnostic and performance metrics tool for routine
502 evaluation of Earth system models in CMIP. *Geoscientific Model Development*, 9: 1747-1802.
- 503 Farquhar, G.D., von Caemmerer, S.v. and Berry, J.A., 1980. A biochemical model of photosynthetic CO₂
504 assimilation in leaves of C₃ species. *Planta*, 149(1): 78-90.
- 505 Fisher, J.B. et al., 2017. The future of evapotranspiration: Global requirements for ecosystem functioning,
506 carbon and climate feedbacks, agricultural management, and water resources. *Water Resources*
507 *Research*, 53(4): 2618-2626.
- 508 Fisher, J.B., Tu, K.P. and Baldocchi, D.D., 2008. Global estimates of the land-atmosphere water flux based on
509 monthly AVHRR and ISLSCP-II data, validated at 16 FLUXNET sites. *Remote Sensing of*
510 *Environment*, 112(3): 901-919.
- 511 Fisher, J.B., Whittaker, R.J. and Malhi, Y., 2011. ET come home: potential evapotranspiration in geographical

512 ecology. *Global Ecology and Biogeography*, 20(1): 1-18.

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

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

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

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

521 Hogue, T.S., Bastidas, L.A., Gupta, H.V. and Sorooshian, S., 2006. Evaluating model performance and
522 parameter behavior for varying levels of land surface model complexity. *Water Resources Research*,
523 42(8).

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

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

528 Jackson, R.D., 1985. Evaluating evapotranspiration at local and regional scales. *Proceedings of the IEEE*, 73(6):
529 1086-1096.

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

532 Jung, M. et al., 2019. The FLUXCOM ensemble of global land-atmosphere energy fluxes. *Scientific data*, 6(1):
533 74.

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

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

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

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

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

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

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

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

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

553 Martens, B. et al., 2018. Towards Estimating Land Evaporation at Field Scales Using GLEAM. *Remote Sensing*,
554 10(11): 1720.

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

556 and globally gridded forcing data. *Geoscientific Model Development*, 9(1): 283-305.

557 Medici, C., Wade, A.J. and Francés, F., 2012. Does increased hydrochemical model complexity decrease
558 robustness? *Journal of Hydrology*, 440-441: 1-13.

559 Miner, G.L., Bauerle, W.L. and Baldocchi, D.D., 2017. Estimating the sensitivity of stomatal conductance to
560 photosynthesis: a review. *Plant, cell & environment*, 40(7): 1214-1238.

561 Mingyue, C., Junbang, W., Shaoqiang, W., Hao, Y. and Yingnian, L., 2019. Temporal and Spatial Distribution of
562 Evapotranspiration and Its Influencing Factors on Qinghai-Tibet Plateau from 1982 to 2014. *Journal of
563 Resources and Ecology*, 10(2): 213-224.

564 Monteith, J.L., 1965. *Evaporation and environment*, Symposia of the society for experimental biology.
565 Cambridge University Press (CUP) Cambridge, pp. 205-234.

566 Mueller, B. et al., 2013. Benchmark products for land evapotranspiration: LandFlux-EVAL multi-data set
567 synthesis.

568 Mueller, B. et al., 2011. Evaluation of global observations-based evapotranspiration datasets and IPCC AR4
569 simulations. *Geophysical Research Letters*, 38(L06402): doi:10.1029/2010GL046230.

570 Oleson, K. et al., 2013. Technical Description of version 4.5 of the Community Land Model (CLM)
571 Coordinating. BOULDER, COLORADO: 80307-3000.

572 Orth, R., Staudinger, M., Seneviratne, S.I., Seibert, J. and Zappa, M., 2015. Does model performance improve
573 with complexity? A case study with three hydrological models. *Journal of Hydrology*, 523: 147-159.

574 Pastorello, G. et al., 2017. A new data set to keep a sharper eye on land-air exchanges. *Eos, Transactions
575 American Geophysical Union (Online)*, 98(8).

576 Peaucelle, M. et al., 2019. Covariations between plant functional traits emerge from constraining
577 parameterization of a terrestrial biosphere model. *Global Ecology and Biogeography*, 28(9):
578 1351-1365.

579 Polhamus, A., Fisher, J.B. and Tu, K.P., 2013. What controls the error structure in evapotranspiration models?
580 *Agricultural and forest meteorology*, 169: 12-24.

581 Priestley, C.H.B. and Taylor, R., 1972. On the assessment of surface heat flux and evaporation using large-scale
582 parameters. *Monthly weather review*, 100(2): 81-92.

583 Randerson, J.T. et al., 2009. Systematic assessment of terrestrial biogeochemistry in coupled climate-carbon
584 models. *Global change biology*, 15(10): 2462-2484.

585 Running, S.W. et al., 2004. A continuous satellite-derived measure of global terrestrial primary production.
586 *Bioscience*, 54(6): 547-560.

587 Schwalm, C.R. et al., 2013. Sensitivity of inferred climate model skill to evaluation decisions: a case study
588 using CMIP5 evapotranspiration. *Environmental Research Letters*, 8(2): 024028.

589 Shan, N. et al., 2019. Modeling canopy conductance and transpiration from solar-induced chlorophyll
590 fluorescence. *Agricultural and Forest Meteorology*, 268: 189-201.

591 Shi, P. et al., 2006. Net ecosystem CO₂ exchange and controlling factors in a steppe—Kobresia meadow on the
592 Tibetan Plateau. *Science in China Series D: Earth Sciences*, 49(S2): 207-218.

593 Shuttleworth, W.J. and Wallace, J., 1985. Evaporation from sparse crops-an energy combination theory.
594 *Quarterly Journal of the Royal Meteorological Society*, 111(469): 839-855.

595 Stofferahn, E. et al., 2019. The Arctic-Boreal vulnerability experiment model benchmarking system.
596 *Environmental Research Letters*, 14(5): 055002.

597 Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. *Journal of
598 Geophysical Research: Atmospheres*, 106(D7): 7183-7192.

599 Vinukollu, R.K., Wood, E.F., Ferguson, C.R. and Fisher, J.B., 2011. Global estimates of evapotranspiration for

600 climate studies using multi-sensor remote sensing data: Evaluation of three process-based approaches.
601 Remote Sensing of Environment, 115: 801-823.

602 Wang, X., Yang, M., Pang, G., Wan, G. and Chen, X., 2014. Simulation and improvement of land surface
603 processes in Nameqie, Central Tibetan Plateau, using the Community Land Model (CLM3.5).
604 Environmental Earth Sciences, 73(11): 7343-7357.

605 Wang, Y. et al., 2017. Changes in mean and extreme temperature and precipitation over the arid region of
606 northwestern China: Observation and projection. Advances in Atmospheric Sciences, 34(3): 289-305.

607 Yang, P., Xia, J., Zhang, Y., Zhan, C. and Qiao, Y., 2018. Comprehensive assessment of drought risk in the arid
608 region of Northwest China based on the global palmer drought severity index gridded data. Science of
609 the Total Environment, 627: 951-962.

610 Yu, G.-R. et al., 2006. Overview of ChinaFLUX and evaluation of its eddy covariance measurement.
611 Agricultural and Forest Meteorology, 137(3-4): 125-137.

612 Zhang, L., Luo, Y., Yu, G. and Zhang, L., 2010. Estimated carbon residence times in three forest ecosystems of
613 eastern China: Applications of probabilistic inversion. Journal of Geophysical Research, 115(G1).

614 Zhang, Y. et al., 2016. Multi-decadal trends in global terrestrial evapotranspiration and its components.
615 Scientific reports, 6: 19124.

616

617 **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

618