1 # Evaluation of water flux predictive models developed using eddy

2 # covariance observations and machine learning: a meta-analysis

3 Haiyang Shi[1,2,4,5], Geping Luo[1,2,3,5], Olaf Hellwich[6], Mingjuan Xie[1,2,4,5], Chen Zhang[1,2], Yu Zhang[1,2], Yuangang

4 Wang[1,2], Xiuliang Yuan[1], Xiaofei Ma[1], Wenqiang Zhang[1,2,4,5], Alishir Kurban[1,2,3,5], Philippe De Maeyer[1,2,4,5] and

5 Tim Van de Voorde[4,5]

7 [1] State Key Laboratory of Desert and Oasis Ecology, Xinjiang Institute of Ecology and Geography, Chinese

8 Academy of Sciences, Urumqi, Xinjiang, 830011, China.

9 [2] University of Chinese Academy of Sciences, 19 (A) Yuquan Road, Beijing, 100049, China.

10 [3] Research Centre for Ecology and Environment of Central Asia, Chinese Academy of Sciences, Urumqi, China.

11 [4] Department of Geography, Ghent University, Ghent 9000, Belgium.

12 [5] Sino-Belgian Joint Laboratory of Geo-Information, Ghent, Belgium and Urumqi, China.

13 [6] Department of Computer Vision & Remote Sensing, Technische Universität Berlin, 10587 Berlin, Germany.

15 *Correspondence to*: **Geping Luo (luogp@ms.xjb.ac.cn) and Olaf Hellwich (olaf.hellwich@tu-berlin.de)**

16 Submitted to *Hydrology and Earth System Sciences*

Abstract.

With the rapid accumulation of water flux observations from global eddy-covariance flux sites, many studies have used data-driven approaches to model water fluxes with various predictors and machine learning algorithms used. However, it is unclear how various model features affect prediction accuracy. To fill this gap, we evaluated this issue based on records of 139 developed models collected from 32 such studies. SVM (average R-squared = 0.82) and RF (average R-squared = 0.81) outperformed over evaluated algorithms with sufficient sample size in both cross-study and intra-study (with the same data) comparisons. The average accuracy of the model applied to arid regions is higher than in other climate types. The average accuracy of the model was slightly lower for forest sites (average R-squared = 0.76) than for croplands and grasslands (average R-squared = 0.8 and 0.79), but higher than for shrubland sites (average R-squared = 0.67). Using Rn/Rs, precipitation, Ta, and FAPAR improved the model accuracy. The combined use of Ta and Rn/Rs is very effective especially in forests, while in grasslands the combination of Ws and Rn/Rs is also effective. Random cross-validation showed higher model accuracy than spatial cross-validation and temporal cross-validation, but spatial cross-validation is more important in spatial extrapolation. The findings of this study are promising to guide future research on such machine learning-based modeling.

1 Introduction

Evapotranspiration (ET) is one of the most important components of the water cycle in terrestrial ecosystems. It also represents the key variable in linking ecosystem functioning, carbon and climate feedback, agricultural management, and water resources (Fisher et al., 2017). The quantification of ET for regions, continents, or the globe can improve our understanding of the water, heat, and carbon interactions, which is important for global change research (Xu et al., 2018). Information on ET has been used in many fields, including, but not limited to, droughts and heatwaves (Miralles et al., 2014), regional water balance closures (Chen et al., 2014; Sahoo et al., 2011), agricultural management (Allen et al., 2011), water resources management (Anderson et al., 2012), biodiversity patterns (Gaston, 2000). In addition, accurate large-scale and long-time series ET prediction at high spatial and temporal resolution has been of great interest (Fisher et al., 2017).

Currently, there are three main approaches for simulation and spatial and temporal prediction of ET: (i) physical models based on remote sensing such as surface energy balance models (Minacapilli et al., 2009; Wagle et al., 2017), Penman-Monteith equation (Mu et al., 2011; Zhang et al., 2010), Priestley-Taylor equation (Miralles et al., 2011); (ii) process-based land surface models, biogeochemical models and hydrological models (Barman et al., 2014; Pan et al., 2015; Sándor et al., 2016; Chen et al., 2019); and (iii) the observation-based machine learning modeling approach with in situ eddy covariance (EC) observations of water flux (Jung et al., 2011; Li et al., 2018; Van Wijk and Bouten, 1999; Xie et al., 2021; Xu et al., 2018; Yang et al., 2006; Zhang et al., 2021). For remote sensing-based physical models and process-based land surface models, some physical processes have not been well characterized due to the lack of understanding of the detailed mechanisms influencing ET under different environmental conditions. For example, the inaccurate representation and estimation of stomatal conductance (Li et al., 2019) and the linearization (McColl, 2020) of the Clausius-Clapeyron relation in the Penman-Monteith equation may introduce both empirical and conceptual errors into estimates of ET. Limited by

55 complicated assumptions and model parametrizations, these process-based models face challenges in the

56 accuracy of their ET estimations over heterogeneous landscapes (Pan et al., 2020; Zhang et al., 2021).

57 Therefore, many researchers have used data-driven approaches for the simulation and prediction of ET with the

58 accumulation of a large volume of measured observational data of water fluxes in the past decades. Various

59 machine learning models have been developed to simulate water fluxes at the flux site scale. Besides, various

60 predictor variables (e.g., meteorological factors, vegetation conditions, and moisture supply conditions) have

61 been incorporated into such models for upscaling (Fang et al., 2020; Jung et al., 2009) of water flux to a larger

62 scale or understanding the driving mechanisms with the variable importance analysis performed in such models.

63

64 However, to date, the systematic assessment of the uncertainty in the processes of water flux prediction models

65 constructed using the machine learning approach is limited. Although considerable effort has been invested in

66 improving the accuracy of such prediction models, our understanding of the expected accuracy of such models

67 under different conditions is still limited. It is still not easy for us to give the general guidelines for selecting

68 appropriate predictor variables and models. Questions such as 'Which predictor variables are the best in water

69 flux simulations?' and 'How to improve the prediction accuracy of water flux effectively?' etc. still confuse the

70 researchers in the field. Therefore, we should synthesize the findings from published studies to determine which

71 predictor variables, machine learning models, and other features can significantly improve the prediction

72 accuracy of water flux. Also, we are interested in understanding under which specific conditions they are more

73 effective.

74

75 A variety of features control the accuracy of such models, including the predictor variables used, the inherent

76 heterogeneity within the dataset, the plant functional type (PFT) and characteristics of the flux sites, model

77 construction/ validation skills, and the algorithms used:

78 a) Predictor variables used: Compared to process-based models, the data used may have a more significant

79   impact on the final model performance in data-driven models. Various biophysical covariates and other

80   environmental factors have been used for the simulation and prediction of water fluxes. The most

81   commonly used factors include mainly precipitation (Prec), air temperature (Ta), wind speed (Ws), net/sun

82   radiation (Rn/Rs), soil temperature (Ts), soil texture, vapor-pressure deficit (VPD), the fraction of absorbed

83   photosynthetically active radiation (FAPAR), vegetation index (e.g., Normalized Difference Vegetation

84   Index (NDVI), Enhanced Vegetation Index (EVI)), Leaf area index (LAI), and carbon fluxes (e.g., Gross

85   Primary Productivity (GPP)). These used predictor variables and their complex interactions drive the

86   fluctuations and variability of water fluxes. They affect the accuracy of water flux simulations in two ways:

87   their actual impact on water fluxes at the process-based level and their spatio-temporal resolution and

88   inherent accuracy. The relationship between water fluxes and these variables at the process-based driving

89   mechanism level is very different under different PFTs, different climate types, and different

90   hydrometeorological conditions. For example, in irrigated croplands in arid regions, water fluxes may be

91   highly correlated with irrigation practices, and thus soil moisture may be a very important predictor

92   variable, and its importance may be significantly higher than in other PFTs. And in models that incorporate

93   data from multiple PFTs, some variables that play important roles in multiple PFTs may have higher

94   importance. In terms of data spatial and temporal resolution, the data for these predictor variables may have

95	different scales. In terms of spatial resolution, meteorological observations such as precipitation and air
96	temperature are at the flux site scale, while factors extracted from satellite remote sensing and reanalysis
97	climate datasets cover a much larger spatial scale (i.e. the grid-scale). This leads to considerable differences
98	in the degree of spatial match between different variables and the site scale EC observations (approximately
99	100 m x 100 m). It is therefore difficult for some variables to be fairly compared in the subsequent
100	importance analysis of driving factors. In terms of temporal resolution, the importance of predictor
101	variables with different temporal resolutions may be variable for models with different time scales (e.g.,
102	half-hourly, daily, and monthly models). For example, the daily or 8-day NDVI data based on MODIS
103	satellite imagery may better capture the temporal dynamics of water fluxes concerning vegetation growth
104	than the 16-daily NDVI data derived from Landsat images. Besides, data on non-temporal dynamic
105	variables such as soil texture cannot explain temporal variability in water fluxes in the data-driven
106	simulations, although soil texture may be important in the interpretation of the actual driving mechanisms
107	of ET (which may need to be quantified in detail in ET simulations by process-based models). In addition,
108	some inherent accuracy issues (e.g., remote sensing-based NDVI may not be effective at high values) of the
109	predictors may propagate into the consequent machine learning models, thus affecting the modeling and our
110	understanding of its importance. Therefore, it is necessary to consider the spatial and temporal resolution of
111	the data and their inherent accuracy for the predictors used in different studies in the systematic evaluation
112	of data-driven water flux simulations.
113	b) The heterogeneity of the dataset and model validation: the volume and inherent spatiotemporal
114	heterogeneity of the training dataset (with more variability and extremes incorporated) may affect model
115	accuracy. Typically, training data with larger regions, multiple sites, multiple PFTs, and longer year spans
116	may have a higher degree of imbalance (Kaur et al., 2019; Van Hulse et al., 2007; Virkkala et al., 2021;
117	Zeng et al., 2020). And in machine learning, in general, modeling with unbalanced data (with significant
118	differences in the distribution between the training and validation sets) may result in lower model accuracy.
119	Currently, the most common ways of model validation include spatial, temporal, and random cross-
120	validation. Spatial validation is mainly to evaluate the ability of the model to be applied in different regions
121	or flux sites with different PFT types, and one of the common methods is 'leave one site out' (Fang et al.,
122	2020; Papale et al., 2015; Zhang et al., 2021). If the data of the site left out for validation differs
123	significantly from the distribution of the training data set, the expected accuracy of the model applied at that
124	site may be low because the trained model may not capture the specific and local relationships between the
125	water flux and the various predictor variables at that site. For temporal validation, to assess the ability of the
126	models to adapt to the interannual variability, typically some years of data are used for training and the
127	remaining years for model validation (Lu and Zhuang, 2010). If a year with extreme climate is used for
128	validation, the accuracy may be low because the training dataset may not contain such extreme climate
129	conditions. In the case of PFTs that are significantly affected by human activities, such as cropland, the
130	possible different crops grown and different land use practices (e.g., irrigation) across years can also lead to
131	low accuracy in temporal validation.
132	c) Various machine learning algorithms: Some machine learning algorithms may have specific advantages
133	when applied to model the relationships between water fluxes and covariates. For example, neural networks
134	may have an advantage in nonlinear fitting, while random forests can avoid serious overfitting problems.

135   However, which algorithm is better overall in different situations (i.e. applied to different data sets)? Which

136   algorithm is generally more accurate than the others when using the same data set? A comprehensive

137   evaluation is important.

138

139   Therefore, to systematically and comprehensively assess the impact of various features in such modeling, we

140   perform a meta-analysis of published water flux simulation studies that combine the flux site water flux

141   observations, various predictors, and machine learning. The accuracy of model records collected from the

142   literature was linked with various model features to assess the impacts of predictor data types, algorithms, and

143   other features on model accuracy. The findings of this study may be promising to improve our understanding of

144   the impact of various features of the models to guide future research on such machine learning-based modeling.

145   **2 Methodology**

146   **2.1 Protocol for selecting the sample of articles**

147   We applied a general query (on December 1st, 2021) on title, abstract, and keywords to include articles with the

148   "OR" operator applied among expressions (Table 1) in the Scopus database. Preferred Reporting Items for

149   Systematic Reviews and Meta-Analyses (PRISMA) (Moher et al., 2009) are followed when filtering the papers.

150   We first excluded articles that obviously did not fit the topic of this study based on the abstract, and then

151   performed the article screening with the full-text reading.

152

153   The inclusion of articles follows the following criteria:

154   a)   Articles were filtered for those with water fluxes (or latent heat) simulated.

155   b)   The water flux or latent heat observations used in the prediction models should be from the eddy-

156   covariance flux measurements.

157   c)   Articles focusing only on gap-filling (Hui et al., 2004) techniques (i.e., the objective was not simulation

158   and extrapolation of water fluxes using machine learning) were excluded.

159   d)   Only articles that used multivariate regression (with the number of covariates greater than or equal to 3)

160   were included.

161   e)   The determination coefficient (R-squared) of the validation step should be reported as the metric of model

162   performance (Shi et al., 2021; Tramontana et al., 2016; Zeng et al., 2020) in the articles.

163   f)   The articles should be published in English-language journals.

164

165   Although RMSE is also often used for model accuracy assessment, its dependence on the magnitude of water

166   flux values makes it difficult to use for fair comparisons between studies. For example, due to the difference in

167   the range of ET values, models developed from flux stations in dry grasslands will typically have lower RMSE

168   than models developed by flux stations based on forests in humid regions. Therefore, RMSE may not be a good

169   metric for cross-study comparisons in this meta-analysis.

170

171   Table 1. Article search: '[A1 OR A2 OR A3...] AND [B1 OR B2 OR B3...] AND [C1 OR C2 OR C3 OR C4...]'

ID	A	B	C
1	Water flux	Eddy covariance	Machine learning
2	Evapotranspiration	Flux tower	Support Vector
3	Latent heat	Flux site	Neural Network
4			Random Forest

172

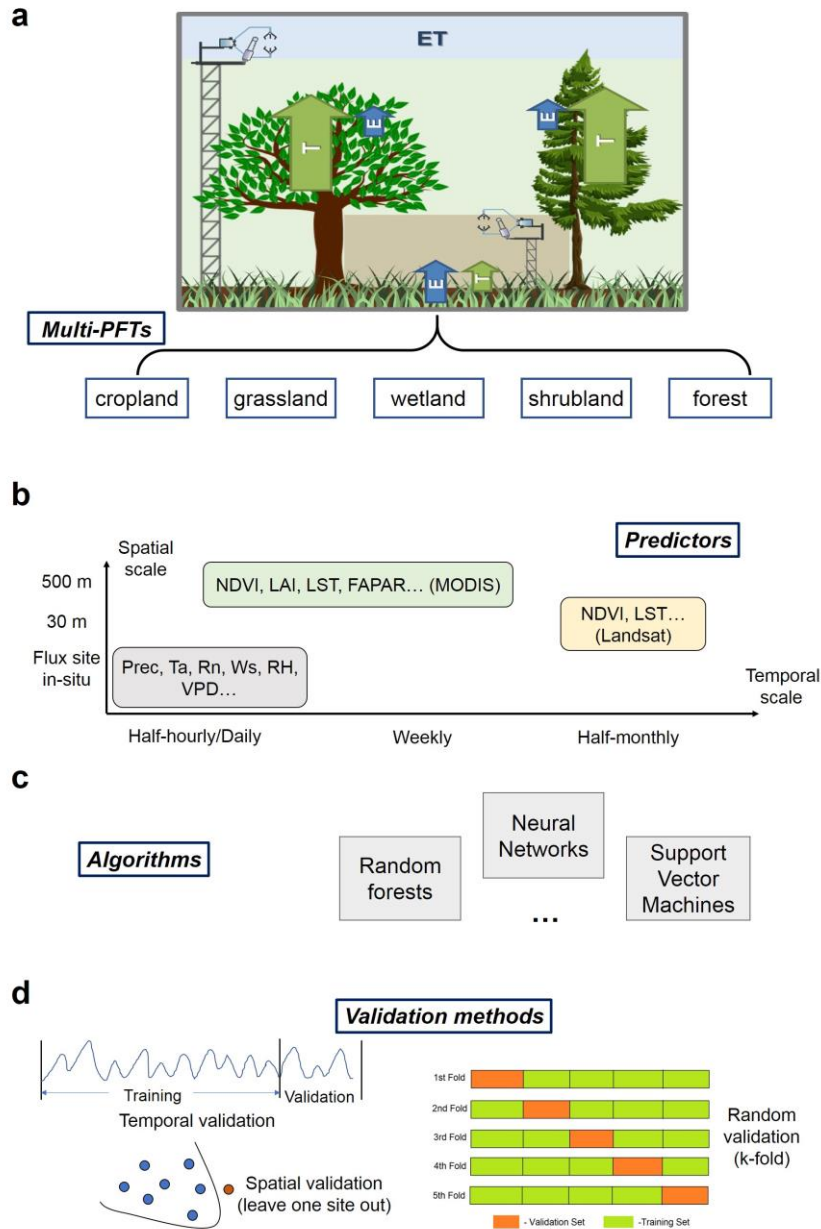
2.2 Features of the prediction processes evaluated

174 The various features (Table 2) involved in the water flux modeling framework (Fig. 1) include the PFTs of the

175 sites, the predictors used, the machine learning algorithms, the validation methods, and other features. Each

176 model for which R-squared is reported is treated as a data record. If multiple algorithms were applied to the

177 same dataset, then multiple records were extracted. Models using different data or features are also recorded as

178 multiple records.

179
180   Figure 1. Features of the machine learning-based water flux prediction process. (a) the eddy-covariance-based
181   water flux observations of various plant function types (PFTs), modified from Paul-Limoges et al., 2020. ET,
182   evapotranspiration. E, evaporation. T, transpiration. (b) Predictors and their spatial and temporal resolution. (c)
183   The machine learning algorithms used for the modeling, such as neural networks, random forests, etc. (d) The
184   model validation methods used including the spatial, temporal, and random cross-validations.

185

186   Table 2. Model feature-related information collected from the papers included in this meta-analysis.

Field	Definition & Categories adopted	Harmonization
Climate	Climate zones of the study location derived from the Köppen climate classification (Peel et al., 2007)	
Plant functional type (PFT)	PFT of the flux sites: 1-forest, 2-grassland, 3-cropland, 4-wetland, 5-shrubland, 6-savannah, and multi-PFTs	The categorization is based on the descriptions in the article. For example, cropland for various crops is classified

		as 'cropland', and both woody savannah and savannah are classified as 'savannah'.
Location	More precise location (with the latitude and longitude of the center of the studied sites): latitude, longitude	
Algorithms	Random Forests (RF), Multiple Linear Regressions (MLR), Artificial Neural Networks (ANN), Support Vector Machines (SVM), Cubist, model tree ensembles (MTE), K-nearest neighbors (KNN), long short-term memory (LSTM), gradient boosting regression tree (GBRT), extra tree regressor (ETR), Gaussian process regression (GPR), Bayesian model averaging (BMA), extreme learning machine (ELM), and deep belief network (DBN)	Various model algorithms with parameter optimization or other improvements are categorized as their algorithm family. For example, various improved models of RF algorithms are classified as RF, rather than as another algorithm family.
Sites number	Number of the flux sites used in the prediction model development	
Spatial scale	Area representatively covered by the flux sites: local (less than 100 x 100 km), regional, global (continent-scale and global scale)	The spatial scale is roughly categorized based on the area covered by the site. The model is classified as 'global' only when the spatial extent reaches the continental scale.
Temporal scale	The temporal scale of the model: half-hourly, hourly, daily, 4-daily, 8-daily, monthly, seasonally (i.e., 0.02, 0.04, 1, 4, 8, 30, 90 days)	Models with a temporal scale greater than one month and less than one year are classified as seasonal scale models.
Year span	The span of years of the flux observations	Year span is calculated as the span from the earliest to the latest year of available flux data.
Site year	The volume of total flux data aggregated by the number of sites and years	
Cross-validation	Strategies in model validation: spatial (e.g., 'leave one site out'), temporal (e.g., 'leave one year out'), random (e.g., 'k-fold')	
Training/validation	The ratio of the data volume in the training and validation sets	In spatial validation, this ratio is represented by the ratio of the number of sites used for training to the number of sites used for validation. In temporal validation, this is represented by the ratio of the span of time periods used for training to the span of time periods used for validation.
Satellite images	The source of satellite images used to derive remote sensing-based predictors (e.g., leaf area index (LAI)), etc: Landsat, MODIS, AVHRR	
Biophysical predictors	NDVI/EVI, the fraction of absorbed photosynthetically active radiation/photosynthetically active radiation (FAPAR/PAR), LAI, Carbon fluxes (CF) including NEE/GPP, etc.	The predictor variables of different measurement methods are categorized according to their definitions. For example, both using the NDVI calculated based on satellite remote

		sensing bands and in situ measurements were classified as the use of 'NDVI'.
Meteorological variables	Precipitation (Prec), net radiation/solar radiation (Rn/Rs), air temperature (Ta), vapour-pressure deficit (VPD), relative humidity (RH) , etc.	The way meteorological data are measured is not differentiated. For example, both using Ta from reanalysis data and Ta measured at flux sites were classified as the use of Ta.
Ancillary data	The ancillary predictor variables used in the modeling: soil texture, terrain (DEM), soil moisture/land surface water index (SM/LSWI), etc	Both the use of in situ measured soil moisture and the use of remote sensing-based LSWI was classified as using surface moisture-related indicators SM/LSWI.
Accuracy metric	Accuracy metric used: R-squared (in the validation phase)	

187

3 Results

3.1 Articles included in the meta-analysis

A total of 32 articles (Table S1) containing a total of 139 model records were included. The geographical scope of these articles was mainly Europe, North America, and China (Fig. 2).

192

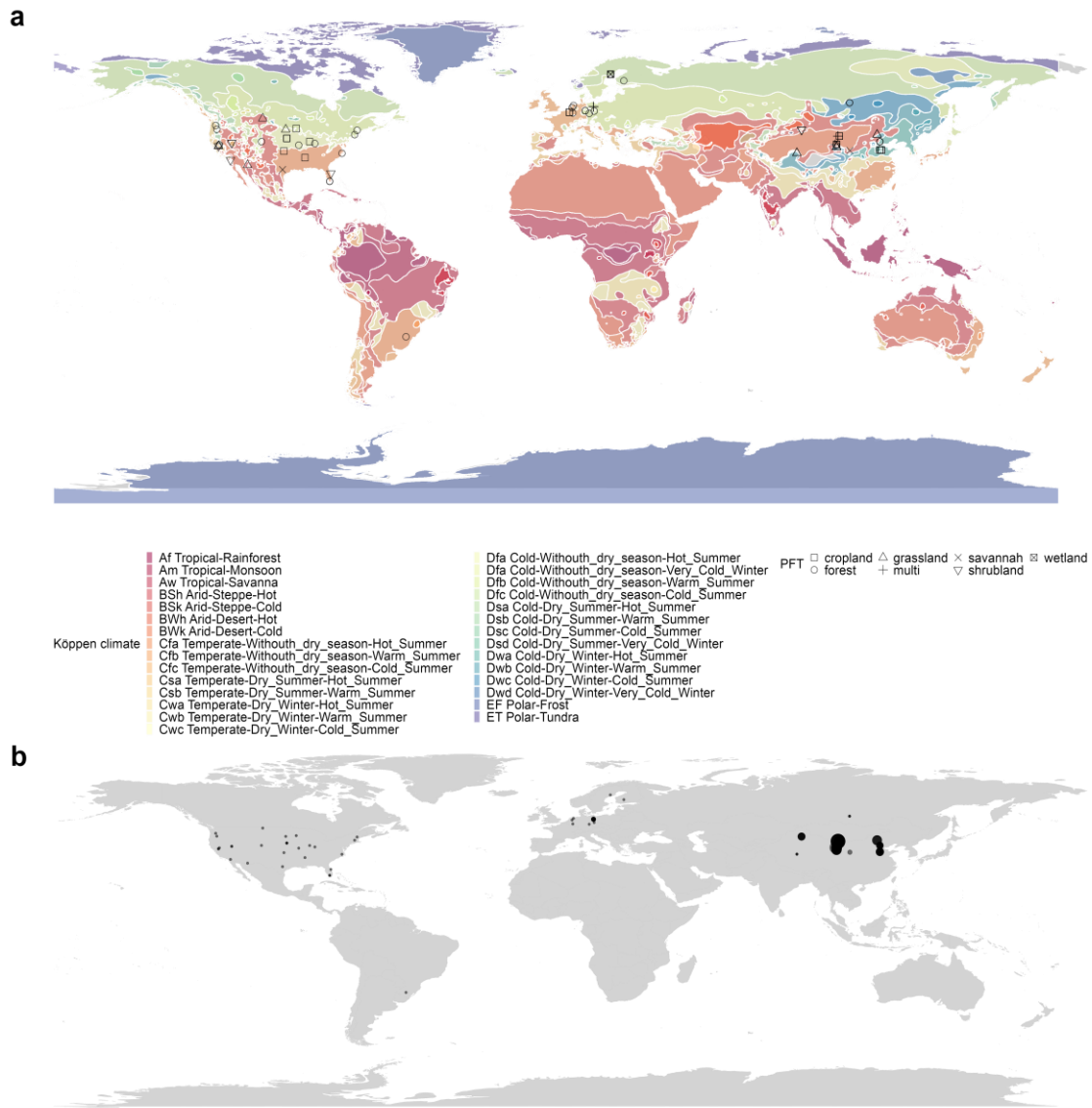



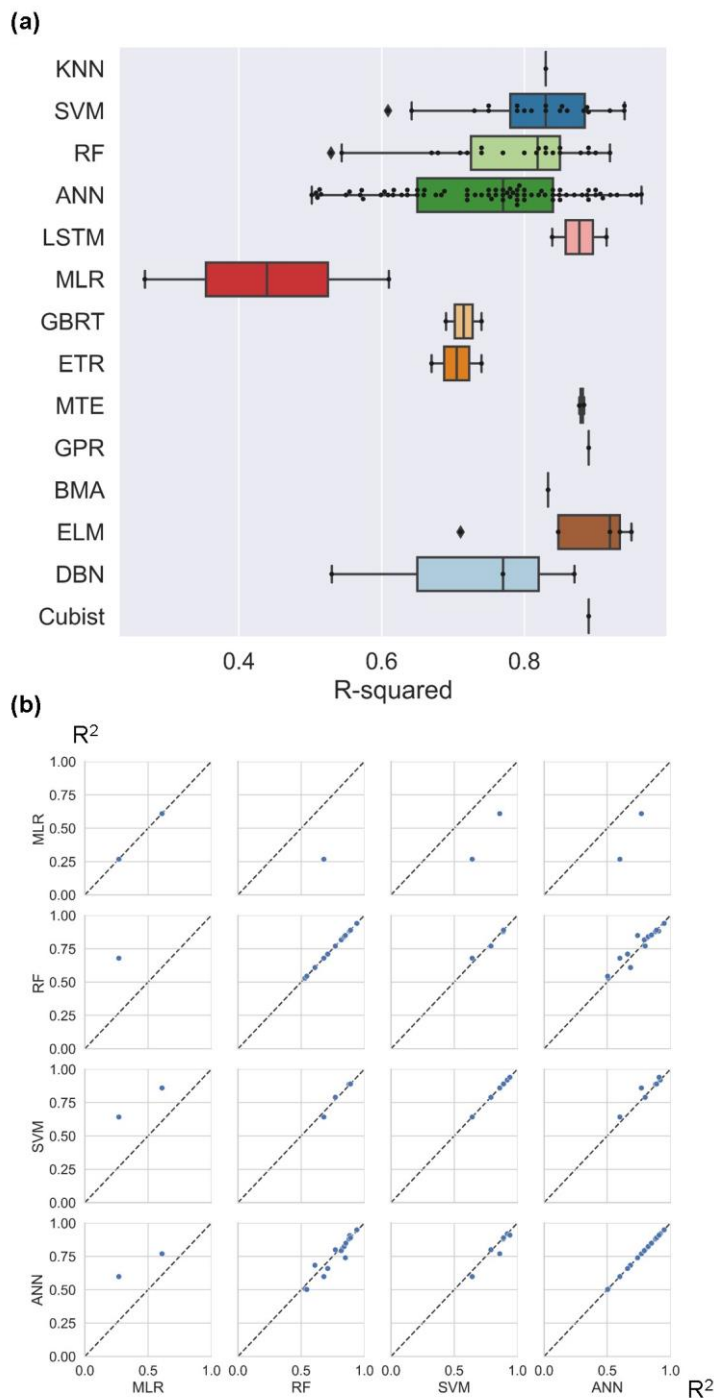
Figure 2. Location of the included studies in the meta-analysis. (a) PFTs and the climate zones (from Köppen climate classification) of these studies and (b) the number of flux sites included in each study. Global and continental-scale studies (e.g., models developed based on FLUXNET of the global scale) are not shown on the map due to the difficulty of identifying specific locations.

3.2 The formal Meta-analysis

3.2.1 Algorithms

SVM and RF outperformed (Fig. 3a) across studies (better than other algorithms with sufficient sample size in Fig. 3a such as ANN). These three machine learning algorithms (i.e., ANN, SVM, RF) were significantly more accurate than the traditional MLR. Other algorithms such as MTE, ELM, Cubist, etc. also correspond to high accuracy, but with limited evidence sample size (Fig. 3a). In the internal comparison (different algorithms applied to the same data set) in single studies, we also find that SVM and RF were slightly more accurate than ANN (Fig. 3b), and all these three (i.e., ANN, SVM, RF) are considerably more accurate than MLR. Overall,

205   SVM and RF have shown higher accuracy in water flux simulations in both inter and intra-study comparisons

206   with sufficient sample size as evidence.

207

208   Figure 3. Model accuracy (R-squared) using various algorithms across studies (a) and internal comparisons of

209   selected pairs of algorithms within studies (b). Algorithms: Random Forests (RF), Multiple Linear Regressions

210   (MLR), Artificial Neural Networks (ANN), Support Vector Machines (SVM), Bayesian model averaging

211   (BMA), Cubist, model tree ensembles (MTE), gradient boosting regression tree (GBRT), extra tree regressor

212   (ETR), K-nearest neighbors (KNN), long short-term memory (LSTM), Gaussian process regression (GPR),

213   extreme learning machine (ELM), and deep belief network (DBN).

3.2.2 Climate types and PFTs

We found higher average model accuracy in arid climate zones (Fig. 4a), such as the Cold semi-arid (steppe) climate (BSk) and Cold desert climate (BWk). Most of these studies were located in northwest China and the western USA. It may be caused by the simpler relationship between water fluxes and biophysical covariates in arid regions. In arid zones, due to the high potential ET, the variability in the actual ET may be largely explained by water availability (moisture supply) and vegetation change with the effect of variability in thermal conditions reduced. As for the various PFTs, the average model accuracy was slightly lower for forest types than for cropland and grassland types (Fig. 4b). The lowest average accuracy was found for shrub sites, which may be related to the difficulty of the remote sensing-based vegetation index (e.g., NDVI) to quantify the physiological and ecological conditions of shrubs (Zeng et al., 2022), and the heterogeneity of the spatial distribution of shrubs within the EC observation area may also cause difficulties in capturing their relationships with biophysical variables. We also found high model accuracy for the wetland type, although records as evidence to support this finding may be limited. Compared to other PFTs, the more steady and adequate water availability in the wetland type may make the variations of water fluxes less explained by other biophysical covariates.

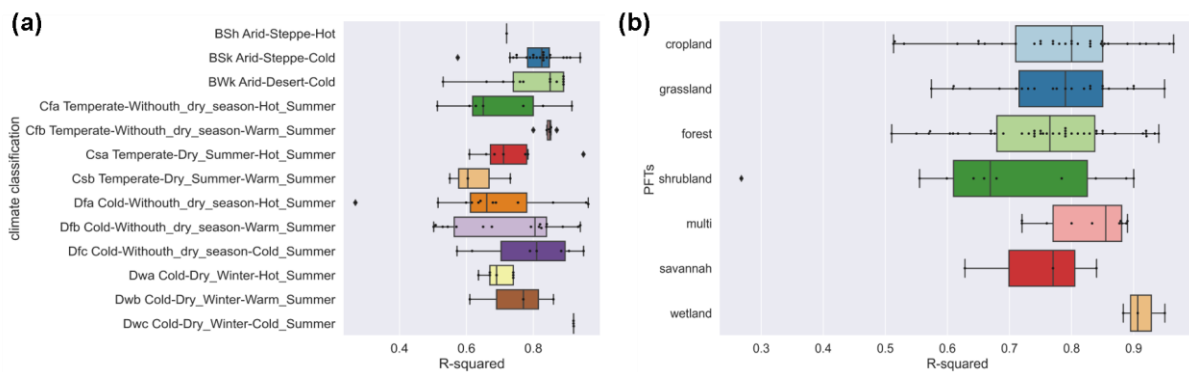



Figure 4. Differences in model accuracy (R-squared) of (a) various climate zones (classified by Köppen climate classification) across studies and (b) PFTs. BSh, Hot semi-arid (steppe) climate. BSk, Cold semi-arid (steppe) climate. BWk, Cold desert climate. Cfa, Humid subtropical climate. Cfb, Temperate oceanic climate. Csa, Hot-summer Mediterranean climate. Csb, Warm-summer Mediterranean climate. Dfa, Hot-summer humid continental climate. Dfb, Warm-summer humid continental climate. Dfc, Subarctic climate. Dwa, Monsoon-influenced hot-summer humid continental climate. Dwb, Monsoon-influenced warm-summer humid continental climate. Dwc, Monsoon-influenced subarctic climate.

3.3.3 Predictors and their combinations

On one hand, for the effects of individual predictors, the use of Rn/Rs, Prec, Ta, and FAPAR improved the accuracy of the model (Fig. S1). This pattern partially changed in the different PFTs. In the forest sites, the accuracy of the models with Rn/Rs and Ta used was higher than that of the models with Rn/Rs and Ta not used. For the grassland sites, the use of Ws, FAPAR, Prec, and Rn/Rs improved the model accuracy. For the cropland sites, Ta and FAPAR were more important for improving the model accuracy.

On the other hand, the evaluation of the effect of individual predictors on model accuracy is not necessarily reliable because some predictor variables are used together (e.g., the high model accuracy corresponding to a particular variable may be because it is often used together with another variable that plays the dominant role in

246   improving accuracy). Therefore, we tested for independence between the use of variables and assessed the effect

247   of the combination of variables on model accuracy. We calculated the correlation matrix (Fig. S2) between the

248   use of various predictors (not used is set as 0 and used is set as 1). We found there was a dependence between

249   the use of some predictors, the use of NDVI/EVI, LAI, and SM was significantly negatively correlated with the

250   use of Rn/Rs and Ta (Fig. S2). It indicated that many of the models that used Rn/Rs and Ta did not use

251   NDVI/EVI, LAI, and SM, and the models that used NDVI/EVI, LAI, and SM also happened to not use Rn/Rs

252   and Ta. Given this dependence, we evaluated the effect of the combination of variables on the model accuracy

253   (Fig. 5). In Fig. 5, the three variable combinations on the left side are mainly meteorological variables while the

254   three variable combinations on the right side are mainly vegetation-related variables based on remote sensing

255   (e.g., NDVI, EVI, LAI, LSWI). We found that, overall, the accuracy of the models using only meteorological

256   variable combinations was higher than that of the models using only remote sensing-based vegetation-related

257   variables. It demonstrated the importance of using meteorological variables in machine learning-based ET

258   prediction (probably especially for models with small time scales such as hourly scale, and daily scale). For

259   example, in the forest type, the combination of Ta and Rn/Rs is very effective compared to using only remote

260   sensing-based vegetation index variable combinations. The combination of Ta and Rn/Rs is also effective in the

261   grassland and cropland types. The combination of Ws and Rn/Rs played an important role in the grassland type

262   for improving model accuracy. Despite this, it does not negate the positive role of remote sensing-based

263   vegetation-related variables in ET prediction. This effectiveness can be dependent on the time scale of the model

264   as well as the PFTs. In models with large time scales (monthly scale, seasonal scale) and PFTs in which ET is

265   sensitive to vegetation dynamics, remote sensing-based vegetation-related variables may also be of high

266   importance.

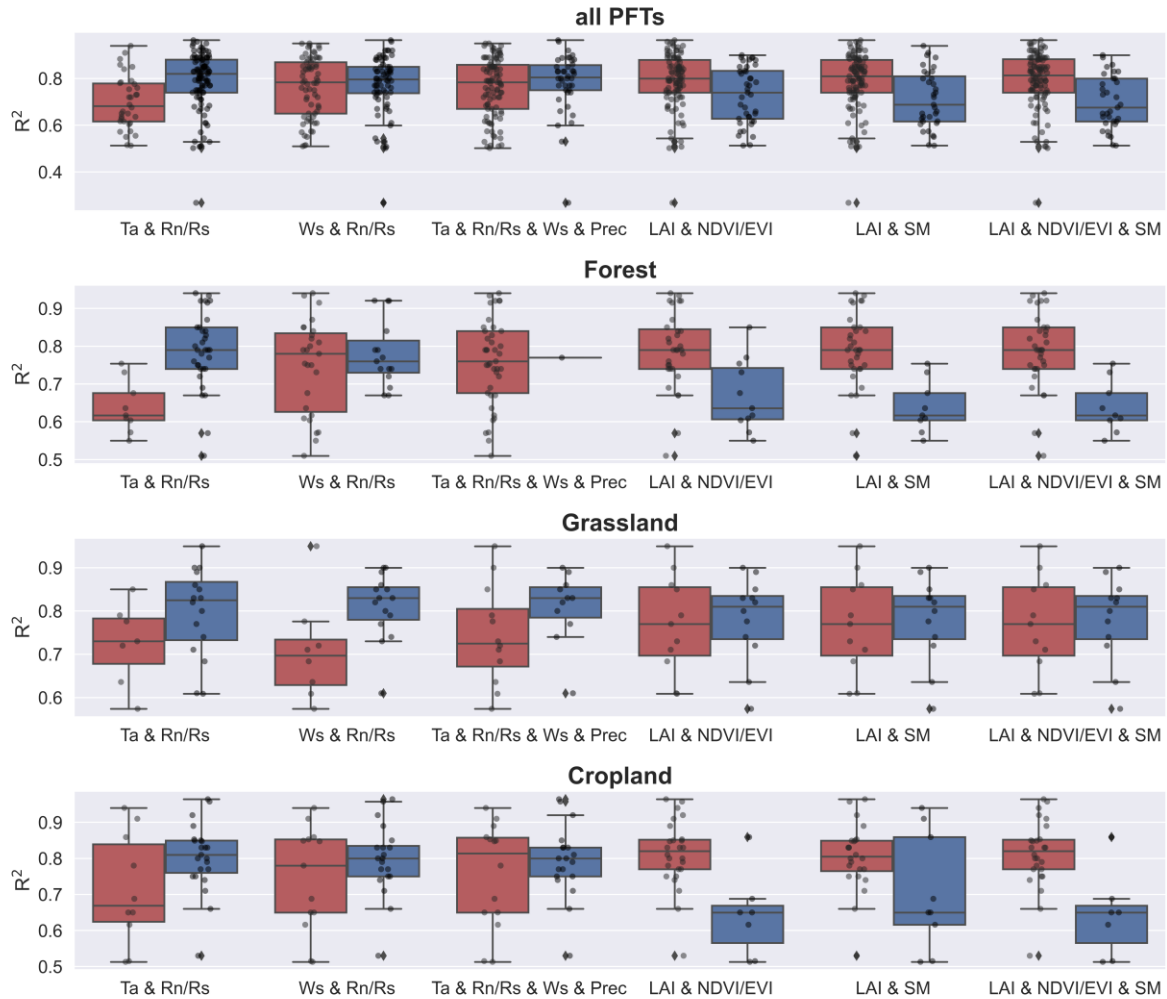



Figure 5. Effects of combinations of predictor variables on model accuracy in various PFTs (all data, forest, grassland, and cropland). Dark blue boxes indicate that the predictors were together used in the model (e.g., for 'Ta & Rn/Rs', the dark blue box represents Ta and Rn/Rs were together used in the model), while dark red boxes indicate the other conditions (i.e., the combination was not used). Predictors: precipitation (Prec), soil moisture/remote sensing-based land surface water index (SM), net radiation/solar radiation (Rn/Rs), enhanced vegetation index (EVI), air temperature (Ta), leaf area index (LAI), Normalized Difference Vegetation Index/Enhanced Vegetation Index (NDVI/EVI).

3.3.4 Other model features

We also evaluated the impact of some other features on accuracy. The differences in the accuracy of models with different spatial scales, year spans, number of sites, and volume of data (Fig. 6) appear to be insignificant. This seems to be related to the fact that in large-scale water flux simulations, the sites of similar PFTs are selected such as for modeling multiple forest sites across Europe (Van Wijk and Bouten, 1999) which focus on 'forest' and multiple grassland sites across arid northern China (Xie et al., 2021; Zhang et al., 2021) which focus on 'grassland', rather than mixing different PFT types to train models as the way in machine learning modeling of carbon fluxes (Zeng et al., 2020). In terms of the time scales of the models, the 4-day, 8-day, and monthly scales appear to correspond to higher accuracy compared to the half-hourly and daily scales. The higher the ratio of the volume of data in the training and validation sets, the higher the model accuracy. Compared to the models

285 using Landsat data, the models using MODIS data showed slightly higher accuracy probably due to the

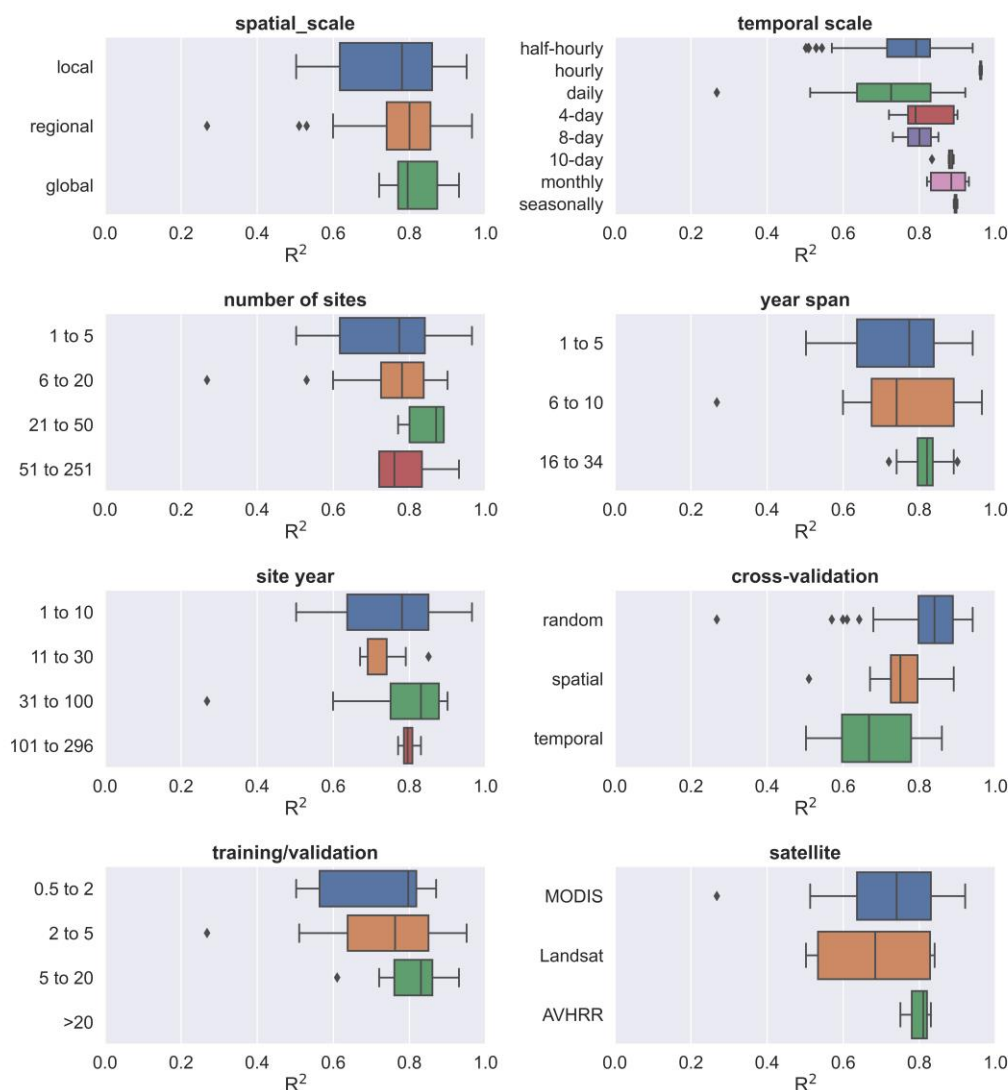
286 advantage of MODIS data in capturing the temporal dynamics of biophysical covariates. There were significant

287 differences in the accuracy of the models using different cross-validation methods, with the models using

288 random cross-validation showing higher accuracy than those using temporal cross-validation. This suggests that

289 interannual variability may have a high impact on the models in water flux simulations. The driving mechanism

290 of ET may vary significantly across years, and the inclusion of some extreme climatic conditions in the training

291 set may be important for model accuracy and robustness.

292

293 Figure 6. The effects of other model features (i.e. spatial scale, number of sites, temporal scale, year span, site

294 year, validation method, training/validation ratio, and satellite imagery used) on the R-squared.

295 **3.3.5 Linear correlation of quantitative features and R-squared**

296 We also analyzed the linear correlation (Fig. 7) between multiple quantitative features and the R-squared. We

297 found that the magnitude of the linear correlation coefficients between the use of predictor combinations and the

298 R-squared was higher than other features. The use of the predictor combination 'Ta and Rn/Rs' significantly

299 improved the model accuracy. 'Temporal scale', 'time span', 'training/validation ratio', and 'number of sites'

showed weak positive correlations with R-squared (not significant, p-value > 0.1). The positive correlation between 'temporal scale' and R-squared is higher among these features, although not significant. It should also be paid more attention to in future studies. The feature 'training/validation ratio' and 'time span' are also positively correlated (although not significantly) with the R-squared, suggesting the importance of the volume of data in the training set in a data-driven machine learning model. Larger 'training/validation ratio' and 'time span' may correspond to greater proportional coverage of the scenarios/conditions in the training set over the validation set, and thus correspond to higher accuracy.

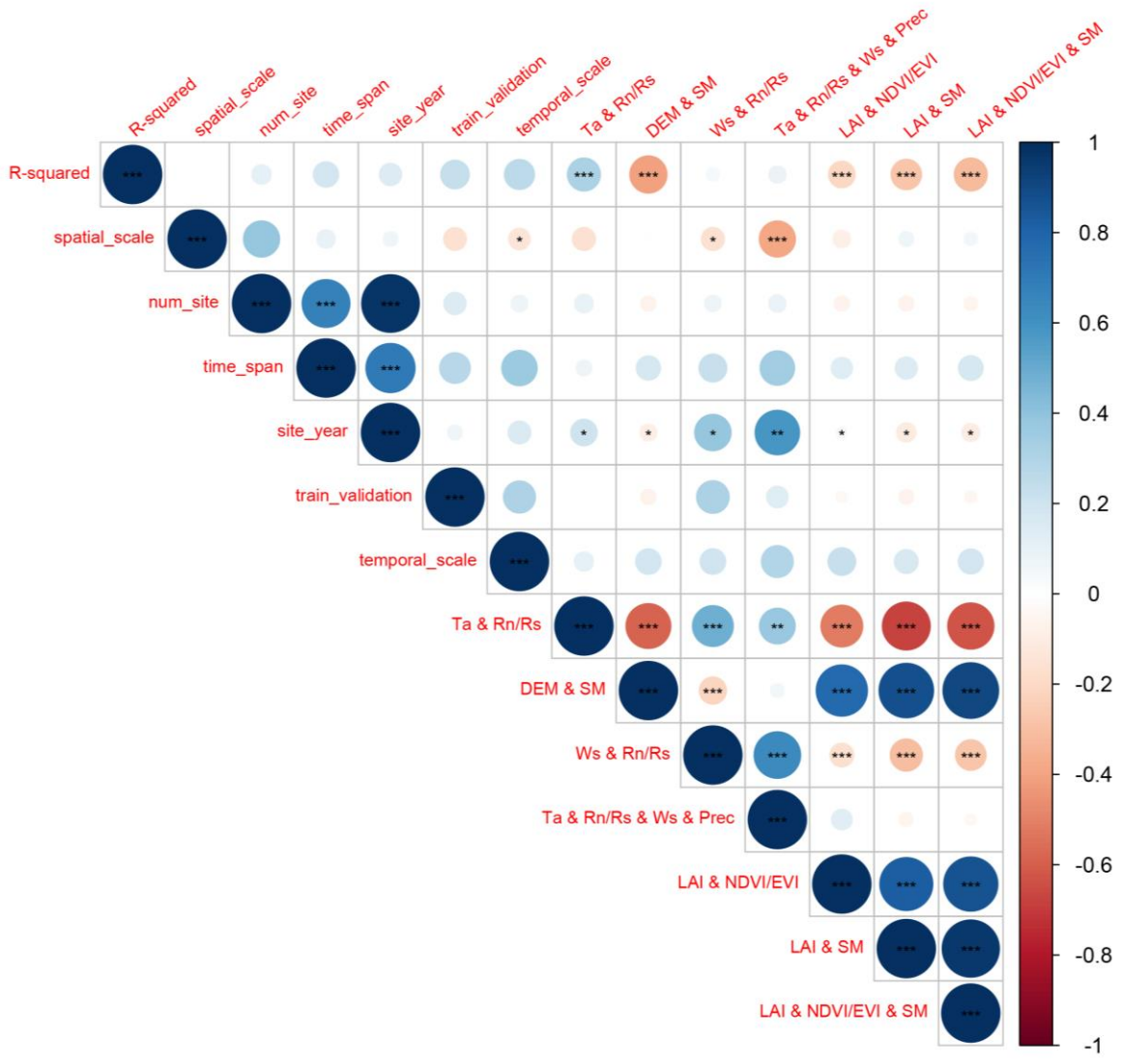



Figure 7. Evaluation of linear correlations between multiple features and the R-squared records with the statistical significance test. For the feature 'spatial scale', the 'local' scale was set to 1, the 'regional' scale was set to 2, and the 'global' scale was set to 3 in the analysis of linear correlation. For the use of various predictor combinations with '&', the value for 'together used' is set as 1 and other conditions are set as 0 (e.g., for the feature 'Ta & Rn/Rs & Ws & Prec', if Ta, Rn/Rs, Ws, and Prec were used together in the model, the value is set as 1). Significance: the p-value < 0.01 (***), 0.05 (**), and 0.1 (*).

4 Discussions

With the accumulation of in situ EC observations around the world, the study of ET simulations based on data-driven approaches has received more attention from researchers in the last decade. Many studies have combined EC observations, various predictors, and machine learning algorithms to improve the prediction accuracy of water fluxes. To date, the results of these studies have not been comprehensively evaluated to provide clear guidance for feature selection in water flux prediction models. To better understand the approach and guide future research, we performed a meta-analysis of such studies. Machine learning-based water flux simulations and predictions still suffer from high uncertainty. By investigating the expected improvements that can be achieved by incorporating different features, we can avoid practices that may reduce model accuracy in future research.

4.1 Opportunities and challenges in the water flux simulation

In the above meta-analysis of the models, we found that water flux simulations based on EC observations can achieve high accuracy but also have high uncertainty through the modeling workflow. The R-squared of many water flux simulation models exceeds 0.8, possibly higher than some remote sensing-based and process-based models, and possibly higher than carbon flux simulations such as the net ecosystem exchange (NEE) in a similar modeling framework (Shi et al., 2022). This may be because many data on important variables affecting carbon flux such as soil and biomass pools, disturbances, ecosystem age, management activities, and land use history are not yet effectively and continuously measured (Jung et al., 2011) with the global spatially and temporally explicit information. While ET simulations rely on observations of moisture and energy conditions and vegetation conditions, much of the current available meteorological and remote sensing data have been effective to represent and capture the spatial and temporal dynamics of these predictors well.

4.1.1 Comprehensive insights on model features

Biophysical and meteorological variables are considered both important in ET simulations. This study found that models using a combination of meteorological variables had higher accuracy than models using only remotely sensed vegetation dynamic information. However, due to the high proportion of models with small temporal scales (e.g., half-hourly scale, hourly scale, and daily scale) in this study, this advantage of the combination of meteorological variables may be more suitable for small temporal scales. A possible explanation is that vegetation-related variables such as NDVI and LAI at the daily scale, 8-day scale, and 16-day scale have limited explanatory ability for hourly or daily-scale variability in ET, especially under cloudy conditions (e.g., tropical rainforest regions), the temporal continuity of the vegetation index data may be greatly limited (Zeng et al., 2022). This should be given more attention and some vegetation indices derived from hourly temporal resolution satellite remote sensing data such as GOES (Zeng et al., 2022) can be used for ET simulations to investigate the possible adding-values of vegetation indices at smaller time scales. In contrast, at a small temporal scale, the use of combinations of meteorological variables can capture moisture and energy conditions that control the rapid fluctuations of ET and thus has a dominant role in hourly or daily-scale ET prediction. This also corroborates the high accuracy of some physic-based ET estimation models (Rigden and Salvucci, 2015) that use only meteorological variables and not vegetation-related variables such NDVI (only an estimate

351 of vegetation height derived from land cover maps is used to represent vegetation conditions (Rigden and

352 Salvucci, 2015)).

353

354 There are differences in model accuracy among different PFTs. For example, in forest sites, limitations in data

355 accuracy of factors were possible because some remote sensing-based predictors such as NDVI, FAPAR, and

356 LAI have limited accuracy when applied to forest types (Liu et al., 2018b; Zeng et al., 2022). In addition, factors

357 such as crown density, which may significantly affect the proportion of soil evaporation, transpiration, and

358 evaporation of canopy interception, were not considered in these models, which may also lead to low model

359 accuracy. This suggests that in water flux simulation, the driving mechanisms of water fluxes in different PFTs

360 do affect the accuracy of machine learning models, and we need to consider more the actual and specific

361 influencing factors in specific PFTs. More variables that can quantify the ratio of evaporation and transpiration

362 should be considered for inclusion, which also appears to improve the mechanistic interpretability of such

363 machine learning models. A previous study (Zhao et al., 2019) combined the physics-based approach (e.g.,

364 Penman-Monteith equation) and machine learning to build hybrid models to improve interpretability. We should

365 make full use of empirical knowledge and experiences from process-based models to improve the accuracy and

366 interpretability of the machine learning approach.

367

368 Among the validation methods, random cross-validation has higher accuracy than spatial cross-validation and

369 temporal cross-validation. However, spatial cross-validation and temporal cross-validation may be able to better

370 help us recognize the robustness of the model when extrapolated (i.e., applied to new stations and new years).

371 The lower accuracy in the temporal cross-validation approach implies that we need to focus on interannual

372 hydrological and meteorological variability in the water flux simulations. In cropland sites, we may also need to

373 pay more attention to the effects of interannual variability in anthropogenic cropping patterns. If some extreme

374 weather years are not included, the robustness of the model when extrapolated to other years may be challenged,

375 especially in the context of the various extreme weather events of recent years. This can also inform the siting of

376 future flux stations. Regions where climate extremes may occur and biogeographic types not covered by

377 existing flux observation networks should be given more attention to achieve global-scale, accurate and robust

378 machine learning-based spatio-temporal prediction of water fluxes. Furthermore, although the R-squared and the

379 training/validation ratio show a positive correlation (Fig. 7) (i.e., a higher training/validation ratio may

380 correspond to a higher R-squared), we should still be cautious in reducing this ratio in our modeling. For a really

381 small validation set, it would be very challenging to determine which model is better given the potential

382 uncertainty caused by the considerable randomness.

383 **4.1.2 Differences from NEE predictions in the similar model framework**

384 In general, predictors related to meteorological, vegetation, and soil conditions were common to both ET and

385 NEE simulations in a similar framework (Shi et al., 2022). However, in NEE predictions, explanatory variables

386 such as soil organic content, photosynthetic photon flux density, and growing degree days (Shi et al., 2022) are

387 not necessary for ET predictions. The selection of these variables requires our prior knowledge of the dominant

388 drivers of ET and NEE anomalies of particular ecosystems and their differences.

389

390 The accuracy of NEE predictions (Shi et al., 2022) can be more limited by global variability across biomes and
391 locations (Nemani et al., 2003) given the lack of locally measured data on soil and biomass pools, disturbances,
392 ecosystem age, management activities and land use history (Jung et al., 2011). It can result in a higher
393 heterogeneity of the training data in large-scale modeling with multiple flux sites (Shi et al., 2022) and the weak
394 ability to capture the NEE anomalies. In contrast, in ET predictions, meteorological variables and vegetation
395 conditions appear to be already sufficient to capture a considerably large fraction of the ET variations in most
396 conditions.
397
398 In future ET prediction studies, given that few current ET products have time scales smaller than daily scale
399 (Jung et al., 2019; Pan et al., 2020), improvements in the accuracy of daily and hourly models may be necessary
400 to fill this gap. Besides, the partitioning of ET components (i.e., transpiration, interception evaporation, and soil
401 evaporation) can be more focused to better decouple the contributions of vegetation and soil to ET with machine
402 learning (Eichelmann et al., 2022). It can be further matched with the partitioning of NEE (i.e., to GPP and
403 ecosystem respiration) to increase our knowledge of the global water cycle and ecosystem functioning and
404 obtain further refined global carbon-water fluxes coupling relations (Eichelmann et al., 2022). Also, the above
405 two promising improvements can be beneficial for research on topics related to the global terrestrial water cycle
406 (Fisher et al., 2017).

407 **4.2 Uncertainties and limitations of this meta-analysis**

408 **4.2.1 The limited number of available literature and model records**
409 Despite many articles and model records collected through our efforts to perform this meta-analysis, there still
410 appears to be a long way to go to finally and completely understand the various mechanisms involved in water
411 flux simulation with machine learning. Some of the insights provided by this study can be not robust (due to the
412 limited sample size available when the goal is to assess the effects of multiple features), but this does not negate
413 the fact that this study does obtain some meaningful findings. Therefore, researchers should treat the results of
414 this study with caution, as they were obtained only statistically. Overall, it is still positive to conduct a meta-
415 analysis of such studies, considering their rapid growth in the number and lack of guiding directions.

416 **4.2.2 Publication bias and weighting**
417 Publication bias and weighting: In a meta-analysis in other fields, weights for different studies can be assigned
418 based on the quality of the journal and the extent to which the research data are publicly available (Borenstein et
419 al., 2011; Field and Gillett, 2010). However, most of the articles included in this study did not fully publish the
420 flux data they used, the models they developed, and the predicted ET data. Given this limitation, we were unable
421 to assign them small weights due to the relatively limited available sample size of this study. Besides, in meta-
422 analyses in other fields, the sample size and the variance of the results of the experiments can also be used to
423 adjust the weights of the effects among studies (Adams et al., 1997; Don et al., 2011; Liu et al., 2018a).
424 However, for this study, due to the lack of a convincing way to determine the weights, we briefly assigned equal
425 weight values to all the included studies.

4.2.3 Uncertainties in the information of the extracted features

At the information extraction level, the following issues may also introduce uncertainties:

a) Uncertainties caused by data quality control (e.g. gap-filling (Hui et al., 2004)) are difficult to assess effectively. Gap-filling is a commonly used technique to fill in low-quality data in flux observations. However, the impact of this practice on machine learning-based ET prediction models is unclear, due to the difficulty of directly assessing how this technique is performed in various studies by this meta-analysis. Typically, models with small time scales (e.g., hourly scale, daily scale) can exclude low-quality observations and use only high-quality data. However, for models with large time scales (e.g., monthly scales), gap-filling (e.g., based on meteorological data) may be unavoidable. This may lead to a decrease in training data purity and introduce uncertainty in the subsequent prediction model development.

b) Systematic uncertainties caused by the energy balance closure (EBC) issue in eddy-covariance flux measurements are also difficult to assess by this meta-analysis. EBC is a common problem (Eshonkulov et al., 2019) in eddy-covariance flux observations. For that reason, the latent heat flux measured potentially underestimates ET. Some prediction models corrected EBC (e.g., using Bowen ratio preserving (Mauder et al., 2013, 2018) and energy balance residuals (Charuchittipan et al., 2014; Mauder et al., 2018)) in the processing of training data, but some did not. How this will affect the accuracy of the prediction model is not clear due to multiple factors that need to be evaluated that influence EBC (Foken, 2008), including measurement errors of the energy balance components, incorrect sensor configurations, influences of heterogeneous canopy height, unconsidered energy storage terms in the soil-plant-atmosphere system, inadequate time averaging intervals, and long-wave eddies (Jacobs et al., 2008; Foken, 2008; Eshonkulov et al., 2019). To reduce this uncertainty, more attention to flux site characteristics (Eshonkulov et al., 2019) related to PFT, topography, flux footprint area, etc., to select the appropriate correction method is necessary for future studies.

c) As most studies used far more water flux observation records than the number of covariates in their regression models, we did not adjust the R-squared in this study to an adjusted R-squared.

d) The various specific ways in which the parameters of the model are optimized are not differentiated. They are broadly categorized into different families or kinds of algorithms, which may also introduce uncertainty into the assessment.

e) The assessment of some features is not detailed due to the limitations of the available model records. For example, the classification of PFT could be more detailed. 'Forest' could be further classified as broadleaf forest, coniferous forest, etc. while 'cropland' could be further classified as rainfed and irrigated cropland based on differences in their response mechanisms of water fluxes to environmental factors.

5 Conclusion

We performed a meta-analysis of the water flux simulation studies which focus on using machine learning approaches to combine flux station observations from flux stations/networks, meteorological, biophysical, and other ancillary predictors. The main conclusions are as follows:

462   1.   SVM (average R-squared = 0.82) and RF (average R-squared = 0.81) outperformed over evaluated
463   algorithms with sufficient sample size in both cross-study and intra-study (with the same training dataset)
464   comparisons.
465   2.   The average accuracy of the model applied to arid regions is higher than in other climate types.
466   3.   The average accuracy of the model was slightly lower for forest sites (average R-squared = 0.76) than for
467   cropland and grassland sites (average R-squared = 0.8 and 0.79), but higher than for shrub sites (average R-
468   squared = 0.67).
469   4.   Among various predictor variables, the use of Rn/Rs, Prec, Ta, and FAPAR improved the model accuracy.
470   The combination of Ta and Rn/Rs is very effective especially in the forest type, while in the grassland type
471   the combination of Ws and Rn/Rs is also effective.
472   5.   Among the different validation methods, random cross-validation shows higher model accuracy than spatial
473   cross-validation and temporal cross-validation.
474
475

476   **Acknowledgements**

477   We thank the editor and two anonymous reviewers for their insightful comments which contributed substantially

478   to the improvement of this manuscript.

479   **Financial support**

480   This research has been supported by the National Natural Science Foundation of China (grant no. U1803243), the

481   Key Projects of the Natural Science Foundation of Xinjiang Autonomous Region (grant no. 2022D01D01), the

482   Strategic Priority Research Program of the Chinese Academy of Sciences (grant no. XDA20060302), and the

483   High-End Foreign Experts project of China.

484   **Author Contributions**

485   HS and GL were responsible for the conceptualization, methodology, formal analysis, investigation, visualization,

486   and writing. OH contributed to the investigation. XM, XY, YW, WZ, MX, CZ and YZ processed the data. AK,

487   TVDV and PDM provided supervision.

488   **Competing interests**

489   The authors declare that they have no conflict of interest.

490   **Code availability**

491   The codes that were used for all analyses are available from the first author (shihaiyang16@mails.ucas.ac.cn)

492   upon request.

493   **Data availability**

494   The data used in this study can be accessed by contacting the first author (shihaiyang16@mails.ucas.ac.cn) upon

495   request.

496

References

Adams, D. C., Gurevitch, J., and Rosenberg, M. S.: Resampling tests for meta‑analysis of ecological data, Ecology, 78, 1277–1283, 1997.

Allen, R. G., Pereira, L. S., Howell, T. A., and Jensen, M. E.: Evapotranspiration information reporting: I. Factors governing measurement accuracy, Agricultural Water Management, 98, 899–920, https://doi.org/10.1016/j.agwat.2010.12.015, 2011.

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

Barman, R., Jain, A. K., and Liang, M.: Climate-driven uncertainties in modeling terrestrial energy and water fluxes: a site-level to global-scale analysis, Global Change Biology, 20, 1885–1900, https://doi.org/10.1111/gcb.12473, 2014.

Borenstein, M., Hedges, L. V., Higgins, J. P., and Rothstein, H. R.: Introduction to meta-analysis, John Wiley & Sons, 2011.

Charuchittipan, D., Babel, W., Mauder, M., Leps, J.-P., and Foken, T.: Extension of the Averaging Time in Eddy-Covariance Measurements and Its Effect on the Energy Balance Closure, Boundary-Layer Meteorol, 152, 303–327, https://doi.org/10.1007/s10546-014-9922-6, 2014.

Chen, Y., Xia, J., Liang, S., Feng, J., Fisher, J. B., Li, X., Li, X., Liu, S., Ma, Z., Miyata, A., Mu, Q., Sun, L., Tang, J., Wang, K., Wen, J., Xue, Y., Yu, G., Zha, T., Zhang, L., Zhang, Q., Zhao, T., Zhao, L., and Yuan, W.: Comparison of satellite-based evapotranspiration models over terrestrial ecosystems in China, Remote Sensing of Environment, 140, 279–293, https://doi.org/10.1016/j.rse.2013.08.045, 2014.

Chen, Y., Wang, S., Ren, Z., Huang, J., Wang, X., Liu, S., Deng, H., and Lin, W.: Increased evapotranspiration from land cover changes intensified water crisis in an arid river basin in northwest China, Journal of Hydrology, 574, 383–397, https://doi.org/10.1016/j.jhydrol.2019.04.045, 2019.

Don, A., Schumacher, J., and Freibauer, A.: Impact of tropical land-use change on soil organic carbon stocks – a meta-analysis, Global Change Biology, 17, 1658–1670, https://doi.org/10.1111/j.1365-2486.2010.02336.x, 2011.

Eichelmann, E., Mantoani, M. C., Chamberlain, S. D., Hemes, K. S., Oikawa, P. Y., Szutu, D., Valach, A., Verfaillie, J., and Baldocchi, D. D.: A novel approach to partitioning evapotranspiration into evaporation and transpiration in flooded ecosystems, Global Change Biology, 28, 990–1007, https://doi.org/10.1111/gcb.15974, 2022.

Eshonkulov, R., Poyda, A., Ingwersen, J., Wizemann, H.-D., Weber, T. K. D., Kremer, P., Högy, P., Pulatov, A., and Streck, T.: Evaluating multi-year, multi-site data on the energy balance closure of eddy-covariance flux measurements at cropland sites in southwestern Germany, Biogeosciences, 16, 521–540, https://doi.org/10.5194/bg-16-521-2019, 2019.

Fang, B., Lei, H., Zhang, Y., Quan, Q., and Yang, D.: Spatio-temporal patterns of evapotranspiration based on upscaling eddy covariance measurements in the dryland of the North China Plain, Agricultural and Forest Meteorology, 281, https://doi.org/10.1016/j.agrformet.2019.107844, 2020.

Field, A. P. and Gillett, R.: How to do a meta‑analysis, British Journal of Mathematical and Statistical Psychology, 63, 665–694, 2010.

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

544 Foken, T.: The energy balance closure problem: An overview, Ecological Applications, 18, 1351–
545 1367, 2008.

546 Gaston, K. J.: Global patterns in biodiversity, Nature, 405, 220–227,
547 https://doi.org/10.1038/35012228, 2000.

548 Hui, D., Wan, S., Su, B., Katul, G., Monson, R., and Luo, Y.: Gap-filling missing data in eddy
549 covariance measurements using multiple imputation (MI) for annual estimations, Agricultural and
550 Forest Meteorology, 121, 93–111, https://doi.org/10.1016/S0168-1923(03)00158-8, 2004.

551 Jacobs, A. F. G., Heusinkveld, B. G., and Holtslag, A. A. M.: Towards Closing the Surface Energy
552 Budget of a Mid-latitude Grassland, Boundary-Layer Meteorol, 126, 125–136,
553 https://doi.org/10.1007/s10546-007-9209-2, 2008.

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

557 Jung, M., Reichstein, M., Margolis, H. A., Cescatti, A., Richardson, A. D., Arain, M. A., Arneth, A.,
558 Bernhofer, C., Bonal, D., Chen, J., Gianelle, D., Gobron, N., Kiely, G., Kutsch, W., Lasslop, G., Law,
559 B. E., Lindroth, A., Merbold, L., Montagnani, L., Moors, E. J., Papale, D., Sottocornola, M., Vaccari,
560 F., and Williams, C.: Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and
561 sensible heat derived from eddy covariance, satellite, and meteorological observations, Journal of
562 Geophysical Research: Biogeosciences, 116, https://doi.org/10.1029/2010JG001566, 2011.

563 Jung, M., Koirala, S., Weber, U., Ichii, K., Gans, F., Camps-Valls, G., Papale, D., Schwalm, C.,
564 Tramontana, G., and Reichstein, M.: The FLUXCOM ensemble of global land-atmosphere energy
565 fluxes, Sci Data, 6, 74, https://doi.org/10.1038/s41597-019-0076-8, 2019.

566 Kaur, H., Pannu, H. S., and Malhi, A. K.: A Systematic Review on Imbalanced Data Challenges in
567 Machine Learning: Applications and Solutions, ACM Comput. Surv., 52, 79:1-79:36,
568 https://doi.org/10.1145/3343440, 2019.

569 Li, X., He, Y., Zeng, Z., Lian, X., Wang, X., Du, M., Jia, G., Li, Y., Ma, Y., Tang, Y., Wang, W., Wu,
570 Z., Yan, J., Yao, Y., Ciais, P., Zhang, X., Zhang, Y., Zhang, Y., Zhou, G., and Piao, S.:
571 Spatiotemporal pattern of terrestrial evapotranspiration in China during the past thirty years,
572 Agricultural and Forest Meteorology, 259, 131–140, https://doi.org/10.1016/j.agrformet.2018.04.020,
573 2018.

574 Li, X., Kang, S., Niu, J., Huo, Z., and Liu, J.: Improving the representation of stomatal responses to
575 CO2 within the Penman–Monteith model to better estimate evapotranspiration responses to climate
576 change, Journal of Hydrology, 572, 692–705, https://doi.org/10.1016/j.jhydrol.2019.03.029, 2019.

577 Liu, Q., Zhang, Y., Liu, B., Amonette, J. E., Lin, Z., Liu, G., Ambus, P., and Xie, Z.: How does
578 biochar influence soil N cycle? A meta-analysis, Plant and soil, 426, 211–225, 2018a.

579 Liu, Y., Xiao, J., Ju, W., Zhu, G., Wu, X., Fan, W., Li, D., and Zhou, Y.: Satellite-derived LAI
580 products exhibit large discrepancies and can lead to substantial uncertainty in simulated carbon and

581   water fluxes, Remote Sensing of Environment, 206, 174–188,
582   https://doi.org/10.1016/j.rse.2017.12.024, 2018b.

583   Lu, X. and Zhuang, Q.: Evaluating evapotranspiration and water-use efficiency of terrestrial
584   ecosystems in the conterminous United States using MODIS and AmeriFlux data, Remote Sensing of
585   Environment, https://doi.org/10.1016/j.rse.2010.04.001, 2010.

586   Mauder, M., Cuntz, M., Drüe, C., Graf, A., Rebmann, C., Schmid, H. P., Schmidt, M., and
587   Steinbrecher, R.: A strategy for quality and uncertainty assessment of long-term eddy-covariance
588   measurements, Agricultural and Forest Meteorology, 169, 122–135,
589   https://doi.org/10.1016/j.agrformet.2012.09.006, 2013.

590   Mauder, M., Genzel, S., Fu, J., Kiese, R., Soltani, M., Steinbrecher, R., Zeeman, M., Banerjee, T., De
591   Roo, F., and Kunstmann, H.: Evaluation of energy balance closure adjustment methods by
592   independent evapotranspiration estimates from lysimeters and hydrological simulations, Hydrological
593   Processes, 32, 39–50, https://doi.org/10.1002/hyp.11397, 2018.

594   McColl, K. A.: Practical and Theoretical Benefits of an Alternative to the Penman-Monteith
595   Evapotranspiration Equation, Water Resources Research, 56, e2020WR027106,
596   https://doi.org/10.1029/2020WR027106, 2020.

597   Minacapilli, M., Agnese, C., Blanda, F., Cammalleri, C., Ciraolo, G., D'Urso, G., Iovino, M., Pumo,
598   D., Provenzano, G., and Rallo, G.: Estimation of actual evapotranspiration of Mediterranean perennial
599   crops by means of remote-sensing based surface energy balance models, Hydrology and Earth System
600   Sciences, 13, 1061–1074, https://doi.org/10.5194/hess-13-1061-2009, 2009.

601   Miralles, D. G., Holmes, T. R. H., De Jeu, R. a. M., Gash, J. H., Meesters, A. G. C. A., and Dolman,
602   A. J.: Global land-surface evaporation estimated from satellite-based observations, Hydrology and
603   Earth System Sciences, 15, 453–469, https://doi.org/10.5194/hess-15-453-2011, 2011.

604   Miralles, D. G., Teuling, A. J., van Heerwaarden, C. C., and Vilà-Guerau de Arellano, J.: Mega-
605   heatwave temperatures due to combined soil desiccation and atmospheric heat accumulation, Nature
606   Geosci, 7, 345–349, https://doi.org/10.1038/ngeo2141, 2014.

607   Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., and Prisma Group: Preferred reporting items for
608   systematic reviews and meta-analyses: the PRISMA statement, PLoS medicine, 6, e1000097, 2009.

609   Mu, Q., Zhao, M., and Running, S. W.: Improvements to a MODIS global terrestrial
610   evapotranspiration algorithm, Remote Sensing of Environment, 115, 1781–1800,
611   https://doi.org/10.1016/j.rse.2011.02.019, 2011.

612   Nemani, R. R., Keeling, C. D., Hashimoto, H., Jolly, W. M., Piper, S. C., Tucker, C. J., Myneni, R.
613   B., and Running, S. W.: Climate-Driven Increases in Global Terrestrial Net Primary Production from
614   1982 to 1999, Science, 300, 1560–1563, https://doi.org/10.1126/science.1082750, 2003.

615   Pan, S., Tian, H., Dangal, S. R. S., Yang, Q., Yang, J., Lu, C., Tao, B., Ren, W., and Ouyang, Z.:
616   Responses of global terrestrial evapotranspiration to climate change and increasing atmospheric CO2
617   in the 21st century, Earth's Future, 3, 15–35, https://doi.org/10.1002/2014EF000263, 2015.

618   Pan, S., Pan, N., Tian, H., Friedlingstein, P., Sitch, S., Shi, H., Arora, V. K., Haverd, V., Jain, A. K.,
619   Kato, E., Lienert, S., Lombardozzi, D., Nabel, J. E. M. S., Ottlé, C., Poulter, B., Zaehle, S., and
620   Running, S. W.: Evaluation of global terrestrial evapotranspiration using state-of-the-art approaches
621   in remote sensing, machine learning and land surface modeling, Hydrology and Earth System
622   Sciences, 24, 1485–1509, https://doi.org/10.5194/hess-24-1485-2020, 2020.

Papale, D., Black, T. A., Carvalhais, N., Cescatti, A., Chen, J., Jung, M., Kiely, G., Lasslop, G., Mahecha, M. D., Margolis, H., Merbold, L., Montagnani, L., Moors, E., Olesen, Jø. E., Reichstein, M., Tramontana, G., Van Gorsel, E., Wohlfahrt, G., and Ráduly, B.: Effect of spatial sampling from European flux towers for estimating carbon and water fluxes with artificial neural networks, Journal of Geophysical Research: Biogeosciences, 120, 1941–1957, https://doi.org/10.1002/2015JG002997, 2015.

Paul-Limoges, E., Wolf, S., Schneider, F. D., Longo, M., Moorcroft, P., Gharun, M., and Damm, A.: Partitioning evapotranspiration with concurrent eddy covariance measurements in a mixed forest, Agricultural and Forest Meteorology, 280, 107786, https://doi.org/10.1016/j.agrformet.2019.107786, 2020.

Peel, M. C., Finlayson, B. L., and McMahon, T. A.: Updated world map of the Köppen-Geiger climate classification, Hydrology and Earth System Sciences, 11, 1633–1644, https://doi.org/10.5194/hess-11-1633-2007, 2007.

Rigden, A. J. and Salvucci, G. D.: Evapotranspiration based on equilibrated relative humidity (ETRHEQ): Evaluation over the continental U.S., Water Resources Research, 51, 2951–2973, https://doi.org/10.1002/2014WR016072, 2015.

Sahoo, A. K., Pan, M., Troy, T. J., Vinukollu, R. K., Sheffield, J., and Wood, E. F.: Reconciling the global terrestrial water budget using satellite remote sensing, Remote Sensing of Environment, 115, 1850–1865, https://doi.org/10.1016/j.rse.2011.03.009, 2011.

Sándor, R., Barcza, Z., Hidy, D., Lellei-Kovács, E., Ma, S., and Bellocchi, G.: Modelling of grassland fluxes in Europe: Evaluation of two biogeochemical models, Agriculture, Ecosystems & Environment, 215, 1–19, https://doi.org/10.1016/j.agee.2015.09.001, 2016.

Shi, H., Hellwich, O., Luo, G., Chen, C., He, H., Ochege, F. U., Van de Voorde, T., Kurban, A., and de Maeyer, P.: A global meta-analysis of soil salinity prediction integrating satellite remote sensing, soil sampling, and machine learning, IEEE Transactions on Geoscience and Remote Sensing, 1–15, https://doi.org/10.1109/TGRS.2021.3109819, 2021.

Shi, H., Luo, G., Hellwich, O., Xie, M., Zhang, C., Zhang, Y., Wang, Y., Yuan, X., Ma, X., and Zhang, W.: Variability and Uncertainty in Flux-Site Scale Net Ecosystem Exchange Simulations Based on Machine Learning and Remote Sensing: A Systematic Evaluation, Biogeosciences Discussions, 1–25, 2022.

Tramontana, G., Jung, M., Schwalm, C. R., Ichii, K., Camps-Valls, G., Ráduly, B., Reichstein, M., Arain, M. A., Cescatti, A., Kiely, G., Merbold, L., Serrano-Ortiz, P., Sickert, S., Wolf, S., and Papale, D.: Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms, Biogeosciences, 13, 4291–4313, https://doi.org/10.5194/bg-13-4291-2016, 2016.

Van Hulse, J., Khoshgoftaar, T. M., and Napolitano, A.: Experimental perspectives on learning from imbalanced data, in: Proceedings of the 24th international conference on Machine learning, New York, NY, USA, 935–942, https://doi.org/10.1145/1273496.1273614, 2007.

Van Wijk, M. T. and Bouten, W.: Water and carbon fluxes above European coniferous forests modelled with artificial neural networks, Ecological Modelling, https://doi.org/10.1016/S0304-3800(99)00101-5, 1999.

Virkkala, A.-M., Aalto, J., Rogers, B. M., Tagesson, T., Treat, C. C., Natali, S. M., Watts, J. D., Potter, S., Lehtonen, A., Mauritz, M., Schuur, E. A. G., Kochendorfer, J., Zona, D., Oechel, W., Kobayashi, H., Humphreys, E., Goeckede, M., Iwata, H., Lafleur, P. M., Euskirchen, E. S., Bokhorst, S., Marushchak, M., Martikainen, P. J., Elberling, B., Voigt, C., Biasi, C., Sonnentag, O., Parmentier,

667 F.-J. W., Ueyama, M., Celis, G., St.Louis, V. L., Emmerton, C. A., Peichl, M., Chi, J., Järveoja, J.,
668 Nilsson, M. B., Oberbauer, S. F., Torn, M. S., Park, S.-J., Dolman, H., Mammarella, I., Chae, N.,
669 Poyatos, R., López-Blanco, E., Christensen, T. R., Kwon, M. J., Sachs, T., Holl, D., and Luoto, M.:
670 Statistical upscaling of ecosystem CO_2 fluxes across the terrestrial tundra and boreal domain:
671 Regional patterns and uncertainties, Global Change Biology, 27, 4040–4059,
672 https://doi.org/10.1111/gcb.15659, 2021.

673 Wagle, P., Bhattarai, N., Gowda, P. H., and Kakani, V. G.: Performance of five surface energy
674 balance models for estimating daily evapotranspiration in high biomass sorghum, ISPRS Journal of
675 Photogrammetry and Remote Sensing, 128, 192–203, https://doi.org/10.1016/j.isprsjprs.2017.03.022,
676 2017.

677 Xie, M., Luo, G., Hellwich, O., Frankl, A., Zhang, W., Chen, C., Zhang, C., and De Maeyer, P.:
678 Simulation of site-scale water fluxes in desert and natural oasis ecosystems of the arid region in
679 Northwest China, Hydrological Processes, 35, e14444, https://doi.org/10.1002/hyp.14444, 2021.

680 Xu, T., Guo, Z., Liu, S., He, X., Meng, Y., Xu, Z., Xia, Y., Xiao, J., Zhang, Y., Ma, Y., and Song, L.:
681 Evaluating Different Machine Learning Methods for Upscaling Evapotranspiration from Flux Towers
682 to the Regional Scale, Journal of Geophysical Research: Atmospheres, 123, 8674–8690,
683 https://doi.org/10.1029/2018JD028447, 2018.

684 Yang, F., White, M. A., Michaelis, A. R., Ichii, K., Hashimoto, H., Votava, P., Zhu, A.-X., and
685 Nemani, R. R.: Prediction of Continental-Scale Evapotranspiration by Combining MODIS and
686 AmeriFlux Data Through Support Vector Machine, IEEE Transactions on Geoscience and Remote
687 Sensing, 44, 3452–3461, https://doi.org/10.1109/TGRS.2006.876297, 2006.

688 Zeng, J., Matsunaga, T., Tan, Z.-H., Saigusa, N., Shirai, T., Tang, Y., Peng, S., and Fukuda, Y.:
689 Global terrestrial carbon fluxes of 1999–2019 estimated by upscaling eddy covariance data with a
690 random forest, Scientific Data, 7, https://doi.org/10.1038/s41597-020-00653-5, 2020.

691 Zeng, Y., Hao, D., Huete, A., Dechant, B., Berry, J., Chen, J. M., Joiner, J., Frankenberg, C., Bond-
692 Lamberty, B., Ryu, Y., Xiao, J., Asrar, G. R., and Chen, M.: Optical vegetation indices for monitoring
693 terrestrial ecosystems globally, Nat Rev Earth Environ, 1–17, https://doi.org/10.1038/s43017-022-
694 00298-5, 2022.

695 Zhang, C., Luo, G., Hellwich, O., Chen, C., Zhang, W., Xie, M., He, H., Shi, H., and Wang, Y.: A
696 framework for estimating actual evapotranspiration at weather stations without flux observations by
697 combining data from MODIS and flux towers through a machine learning approach, Journal of
698 Hydrology, 603, 127047, https://doi.org/10.1016/j.jhydrol.2021.127047, 2021.

699 Zhang, K., Kimball, J. S., Nemani, R. R., and Running, S. W.: A continuous satellite-derived global
700 record of land surface evapotranspiration from 1983 to 2006, Water Resources Research, 46,
701 https://doi.org/10.1029/2009WR008800, 2010.

702 Zhao, W. L., Gentine, P., Reichstein, M., Zhang, Y., Zhou, S., Wen, Y., Lin, C., Li, X., and Qiu, G.
703 Y.: Physics-Constrained Machine Learning of Evapotranspiration, Geophysical Research Letters, 46,
704 14496–14507, https://doi.org/10.1029/2019GL085291, 2019.

705
706