https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

A Novel Classifier-Guided Ensemble Framework for Global Terrestrial Evapotranspiration Estimates

Le Ni [1,2], Weiguang Wang[1, 2, 3*], Jianyu Fu [1,4], Mingzhu Cao[1,4]

[1]State Key Laboratory of Water Disaster Prevention, Hohai University, Nanjing 210098, China.
[2]College of Hydrology and Water Resources, Hohai University, Nanjing 210098, China.
[3]Yangtze Institute for Conservation and Development, Hohai University, Nanjing 210098, China.
[4]School of Civil Engineering, Sun Yat-sen University, Guangzhou 519082, China

Correspondence to: Weiguang Wang (wangweiguang006@126.com)

Abstract. Evapotranspiration (ET) is a key hydrological and meteorological variable, serving as the critical nexus between water and energy exchanges. However, accurate estimation of global ET remains a challenging task, as process-based ET algorithms are often inadequate to capture the nonlinear relationship among environmental factors, and the application of data-driven ET algorithms is hindered by sparse and uncertain ET observations. In this study, we developed a novel ensemble framework that integrates three existing ET models (process-based algorithm, machine learning-based ET model, and hybrid model), aiming to provide high-precision terrestrial ET estimates. The framework is guided by an additional classifier that can achieve dynamic per-pixel model selection, thus fully utilizing the spatiotemporal dynamics of each model's distinct advantages in mapping global ET and avoiding the typical underestimation of high values by ensemble methods. Comprehensive validation of the model was carried out using in-situ ET observations from the FLUXNET2015 dataset, catchment-scale water balance ET dataset, and six global-scale ET products, including comparisons to individual base models and another Attention-Based ensemble model. The quantitative comparisons across statistical metrics (RMSE, MAE, R^2, KGE) indicate that our ensemble model outperforms other evaluated models, especially in extreme samples. Meanwhile, the introduction of classifier can not only significantly enhance the algorithmic robustness and generalizability, but also allow us to gain a basic understanding of the mechanisms behind model selection by interpretability analysis. The study demonstrated the effectiveness of the proposed framework in enhancing ET estimation robustness, thereby providing a valuable reference for the estimation of other similar variables.

1 Introduction

Evapotranspiration (ET) plays a crucial role in both global water and energy cycles (Fisher et al., 2017; Good et al., 2015; Milly et al., 2005), transferring over 60% of terrestrial precipitation back into the atmosphere (Oki & Kanae, 2006), and concurrently consuming a significant amount of energy (Trenberth et al., 2009). Particularly in the context of global warming, changes in ET will not only alter the distribution of global available freshwater resources (Greve & Seneviratne, 2015; Huntington, 2006; Purdy et al., 2018), but also significantly impact the frequency and severity of hydroclimatic extremes (Miralles et al., 2019; Schwalm et al., 2017). Therefore, reliable ET monitoring at the global scale is of great

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

importance in studying the potential changes in the water cycle and energy budget under climate change conditions (Jung et al., 2011; Milly et al., 2005; Wang & Alimohammadi, 2012; Wang & Dickinson, 2012; Xie et al., 2015; Zhang et al., 2019).

Flux towers can provide reliable in situ ET observations at hourly to sub-hourly timesteps based on the eddy covariance method (Williams et al., 2004; Wilson et al., 2001), but their limited spatial representativeness hinders the acquisition of regional ET. Although regional ET can be measured indirectly through assessing the water balance of catchments, the method is only suitable for catchment ET measurements over annual or longer timescales (Reitz et al., 2023). As none of existing methods can provide direct global ET measurements with both high precision and continuous spatial coverage (Fisher et al., 2017; Reitz et al., 2023; Teuling et al., 2009), remote sensing ET algorithms tend to be used to quantify global ET based on temporally and spatially continuous satellite data (e.g., Bastiaanssen et al., 1998; Jung et al., 2011; Kustas & Norman, 1997; Mu et al., 2011).

The existing remote sensing ET algorithms can be divided into two categories: process-based algorithms and data-driven algorithms (Fu et al., 2022; Shang et al., 2023). Process-based algorithms (e.g., Monteith, 1965; Penman, 1948; Priestley & Taylor, 1972; Su, 2002) employ flux equations to estimate ET based on physically-founded methods, such as the Monin-Obukhov similarity theory, energy balance method and aerodynamic method (Monteith, 1965; Penman, 1948; Allen et al., 1998). However, uncertainties remain in process-based ET algorithms, arising from the insufficient theoretical bases on the complex physical and biological factors involved in ET processes (Mu et al., 2011; Polhamus et al., 2013). With the influx of satellite and in situ observations, data-driven algorithms, especially the machine learning (ML) methods, have become popular in large-scale ET estimation (e.g., Xu et al., 2018; Zhang et al., 2022; Granata, 2019; Lyu & Yong, 2024). ML-based ET models can characterize the nonlinear relationship between different ET-related variables and efficiently capture the spatiotemporal dynamics features of ET from meteorological data streams (Reichstein et al., 2019), thus providing overall more accurate ET estimation than process-based algorithms in data-dense regions. However, due to the lack of global-scale ET observations, these ML-based ET models have to be trained based on in situ observations (Jung et al., 2010). The density of in situ observations is insufficient to represent global ET information, particularly in heterogeneous and data-sparse regions, hindering the use of these ML-based ET models at the global scale (Zhao et al., 2019).

Combining ML-based ET models with process-based algorithms may be a feasible way to improve the generalizability of ML-based ET models (e.g., Brenowitz & Bretherton, 2018; Karpatne et al., 2017; Reichstein et al., 2019; Willard et al., 2022). In these hybrid models, ML methods can be employed for improving parameterizations, or replacing a sub-model of physical model (Reichstein et al., 2019). For example, ML models can be used to estimate the parameters with high uncertainty, such as surface resistance (r_s) in PM equation (Chen et al., 2022; Shang et al., 2023), or to estimate both aerodynamic resistance (r_a) and r_s (ElGhawi et al., 2023). ML models can also be coupled to stress-based ET models by replacing the formulation of transpiration stress (S_t) (Koppa et al. 2022). However, hybrid models still rely on sufficient availability of training data; therefore, they cannot take the place of process-based algorithms, especially in data-sparse regions and heterogeneous surfaces (Shang et al., 2023). On the other hand, due to the uncertainties in the coupling of machine learning and physical laws (Shang et al., 2023), hybrid models cannot consistently outperform pure ML-based ET

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

models in the data-dense regions.

Given the different characteristics of process-based algorithms, ML-based ET models, and hybrid models, it is essential to explore a method to utilize the distinct advantages of the three models. Ensemble learning, a common approach to integrate multiple ML models to achieve better performance (Ganaie et al., 2022; Mohammed & Kora, 2023), may have the potential to address this issue. Several existing ensemble frameworks have been demonstrated to have the capability to achieve better performance than single model (e.g., Pérez-Rodríguez et al., 2023; Tseng, 2023). For example, genetic algorithm can be employed to determine the weights for the ensemble of multiple different ML models (Ayan et al. 2020). Similarly, the attention mechanism in neural network can dynamically assign weights to provide effective model integration (Liu et al. 2022). Statistical methods, such as Bayesian model averaging, can utilize the probability distributions of each model to assess their relative prediction performance, thereby assigning ensemble weights (Huang & Merwade, 2023). However, these approaches exhibit limitations in global ET estimation, either due to the sparse distribution of in situ observations, similar to the challenges encountered by pure ML-based ET models, or due to the non-dynamic weight assignments that cannot reflect the spatiotemporal distribution of distinct model advantages. Previous studies have demonstrated the significantly superior performance of observation-calibrated ML-based models over process-based algorithms at the site scale (Shang et al., 2023), thus when only site-observed ET is the most reliable data source, existing data-driven ensemble methods may not fully utilize the advantages of process-based algorithms. In addition, the existing ensemble models mainly focus on the integration between ML models and whether they are efficient to the integration among ML models, process-based algorithms and hybrid models have not been substantially validated.

Hence, we proposed a novel ensemble framework and developed a model called Classifier-Guided Ensemble model to utilize the individual advantages of three base models (process-based algorithms, ML-based ET models, and hybrid models) by decomposing the ET estimation process into two steps, that is, the classification of input data and the regression of ET. An additional explainable ML classifier was trained to dynamically select the 'dominant model' to be used at each pixel. Since the ML classifier is used for classification rather than directly calculating ET, both in situ ET observations and global ET datasets can serve as reference datasets for deriving classifier training labels, resulting in improved classification accuracy and generalizability of the ensemble framework. In this study, the main objectives are to (a) use the proposed ensemble framework to generate global ET estimation based on in situ observations, satellite retrievals, reanalysis data, and multiple ET products; (b) carry out comprehensive evaluation of the model across multiple spatial scales to analyze model's robustness and generalizability; (c) assess the impact of introducing ML classifier; (d) analyze the interpretability of the ML classifier to gain insights into the implicit meteorological and vegetation features suitable for different ET models. In doing so, our framework offers a reference for ET estimation and contributes to the understanding of the mechanisms behind ET estimation.

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

2 Methodology

The same input covariates for all ML model used were selected: International Geosphere-Biosphere Programme (IGBP) land cover types, leaf area index (LAI), normalized difference vegetation index (NDVI), atmospheric pressure (P), incident solar radiation (Rs), soil moisture (SM), air temperature (Ta), soil temperature (Ts), vapor pressure deficit (VPD), wind speed (WS), with a monthly temporal scale, because these variables are key parameters in ET mechanisms and have been proved to be effective for ET estimation in other studies (Koppa et al., 2022; Shang et al., 2023). The calculation process and the models used are as follows:

2.1. Machine learning model

We chose to use Autogluon for all machine learning components in this study. Autogluon is an open-source AutoML framework that can automatically conduct the selection, combination, and parameterization of multiple ML methods, allowing us to achieve high-accuracy results without manual intervention (Erickson et al., 2020).

Several ML algorithms are provided by Autogluon, including k-Nearest Neighbors, Extremely Randomized Trees, LightGBM boosted trees (Ke et al., 2017), CatBoost boosted trees (Dorogush et al., 2018), Random Forests (Breiman, 2001), neural networks, etc. These models have been widely used with their own distinct characteristics and advantages (Fan et al., 2019; da Silva Júnior et al., 2019; Zhangzhong et al., 2023). Autogluon can combine them using methods known as stacking and bagging (Erickson et al., 2020), and can achieve better performance than individual models. More detailed algorithm information can be found in Erickson et al. (2020).

2.2. Hybrid model

The original P-M equation (Monteith, 1965; Penman, 1948) is as follows:

$$\lambda E_{PM} = \frac{\Delta(R_n - G) + \rho \cdot C_P \cdot VPD / r_a}{\Delta + \gamma \cdot (1 + r_s / r_a)} \quad (1)$$

where λE_{PM} is the latent heat flux (W m^{-2}), Δ is the slope of the saturated vapor pressure vs temperature curve (k Pa $^{\circ}$C^{-1}), R_n is the net radiation (W m^{-2}), G is the soil heat flux (W m^{-2}), ρ is the air density (kg m^{-3}), C_p is the specific heat capacity of air at constant pressure (J kg^{-1} k^{-1}), VPD is the vapor pressure deficit of the air (Pa), γ is the psychrometric constant (k Pa $^{\circ}$C^{-1}), r_a and r_s are the aerodynamic resistance and surface resistance (s m^{-1}).

Although some studies have optimized the estimation of parameter r_s (Wang et al., 2010a, 2010b), estimating parameter r_s remains a challenging task. So in hybrid model, we replaced the empirical expression of r_s with ML model, similar to the surface conductance-based ML model as proposed by Shang et al. (2023). The target label r_s in ML model is obtained by inverting the Equation 1, due to the lack of observations for parameter r_s. As the other variables in Equation 1 can be calculated based on the covariates for ML model, the estimated ET can be computed by Equation 1 after obtaining the parameter r_s from the ML model.

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

2.3. Process-based ET algorithms

We chose to use a MODIS global terrestrial ET algorithm from Mu et al. (2011, 2007) based on the PM equation. They improved the methods to estimate some parameters in traditional PM algorithms and included additional ET sources into the algorithm. They divided ET into three main components: wet canopy evaporation, plant transpiration, and soil evaporation, with soil evaporation further divided into the saturated surface and the moist surface. Some of the main formulas are listed below:

$$\lambda E = \lambda E_{wet_c} + \lambda E_{trans} + \lambda E_{soil} \tag{2}$$

$$\lambda E_{\mathrm{wet_c}} = \frac{\left(s\times A_C\times F_C+\rho\times C_p\times(e_{sat}-e)\times\frac{F_C}{rhrc}\right)\times Fwet}{s+\frac{P_a\times C_p\times rvc}{\lambda\times\varepsilon\times rhrc}} \tag{3}$$

$$\lambda E_{trans} = \frac{\left(s\times A_C\times F_C+\rho\times C_p\times(e_{sat}-e)\times\frac{F_C}{r_a}\right)\times(1-Fwet)}{s+\gamma\times\left(1+\frac{r_s}{r_a}\right)} \tag{4}$$

$$\lambda E_{wet_{soil}} = \frac{\left(s\times A_{soil}+\rho\times C_p\times(1.0-F_C)\times\frac{VPD}{r_{as}}\right)\times Fwet}{s+\gamma\times\frac{r_{\mathrm{tot}}}{r_{as}}} \tag{5}$$

$$\lambda E_{soil_{pot}} = \frac{\left(s\times A_{SOIL}+\rho\times C_p\times(1.0-F_C)\times\frac{VPD}{r_{as}}\right)\times(1.0-Fwet)}{s+\gamma\times\frac{r_{tot}}{r_{as}}} \tag{6}$$

$$\lambda E_{SOIL} = \lambda E_{wet_{soil}} + \lambda E_{soil_{pot}} \times \left(\frac{RH}{100}\right)^{\frac{VPD}{\beta}} \tag{7}$$

where $\lambda E_{\mathrm{wet\_c}}$ is evaporation from wet canopy surface, λE_{trans} is plant transpiration, λE_{wet_soil} and λE_{soil_pot} are evaporation from soil surface and potential soil evaporation, respectively.

Additionally, they improved the method to estimate vegetation cover fraction, soil heat flux, and parameters such as r_s, r_a, etc., and calculated ET as the sum of daytime and nighttime components, thereby enhancing accuracy. More detailed information on the algorithms and the parameters can be found in Mu et al. (2011, 2007). In applying this algorithm, we mainly used input variables LAI, IGBP, P, Rs, Ta, VPD, etc., all of which were also employed by the ML models, without introducing any additional data.

2.4 Classifier-Guided Ensemble model

The Classifier-Guided Ensemble model aims to integrate three base models (process-based algorithm, ML-based ET model, and Hybrid model) to optimize the global ET estimation by dividing the ET estimation process into two sub - problems: (a) training a ML classifier to identify the ‘dominant model’ at each pixel, and (b) using the corresponding ‘dominant model’ at each pixel to estimate ET (Fig. 1).

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

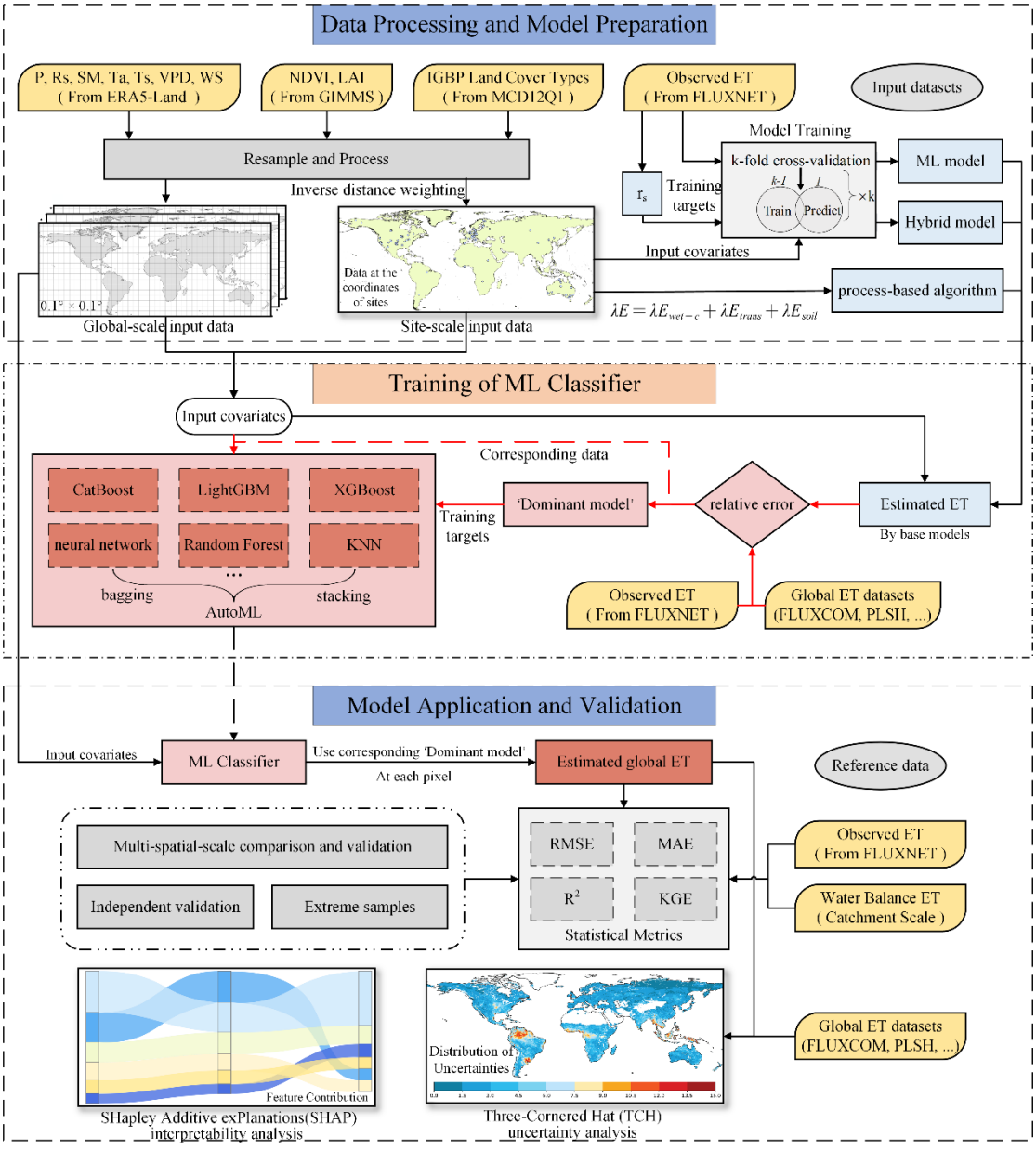

Figure 1. Schematic of the Classifier-Guided Ensemble model. The red arrows indicate the modeling steps of the ML Classifier. P is atmospheric pressure, Rs is incident solar radiation, SM is soil moisture, Ta is air temperature, Ts is soil temperature, VPD is vapor pressure deficit, WS is wind speed, LAI is leaf area index, NDVI is normalized difference vegetation index.

Before training the ML classifier, we needed to first use the three base models to estimate ET at both site and global scales. The estimated ET, in situ ET observations and other global ET products were processed for classification task, in order to obtain the training target. At the site scale, we classified the data at each site for every time point into three types ('ML model-dominated', 'hybrid model-dominated', and 'process-based algorithm-dominated') based on the relative errors

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

between the model-estimated ET and the observed ET. The model with the smallest relative error was the 'dominant model' for the corresponding site and time. Specifically, at the site scale, we conducted ten-fold cross-validation on three base models and used ET estimates from the validation sets for classification task, to avoid abnormally high accuracy resulting from model overfitting.

At the global scale, due to the lack of reliable ET observations, we used six widely used global ET products as references to extract some relatively reliable data from the global dataset for the classification task. We calculated the relative errors between estimated ET of base models and global ET products at each pixel for every time point, obtaining six error values for each base model. If all six relative error values of a base model were lower than those of the other two base models, data of this pixel and time point was added to the training data and the model was considered to be the 'dominant model' of this pixel and time point. Due to differences in the spatial patterns of various ET products, data from other pixels were excluded to reduce uncertainty.

In the training set, we performed the aforementioned classification task, using the classification results as training target for the ML classifier, with the other data serving as input covariates. After training the ML classifier, we could use the global covariate dataset in the validation set to obtain classification results for each pixel and time point, and employ the corresponding 'dominant model' to produce global ET estimates.

2.5 Other ensemble models used for comparison

We use the Attention-Based ensemble technique used by Liu et al. (2022) as a comparison. The Attention Mechanism is similar to human selective attention and is a method to mimic the human visual and cognitive systems. This technique utilizes the focusing ability of the self-attention mechanism (Zhang et al., 2021), which allows the neural network to focus on what it considers important. It can improve model performance by automatically assigning higher weights to sub-models with higher accuracy. The core formula of this model is as follows:

$$\begin{aligned} f_{AE}(x) &= \sum_{i=1}^{N} Attention_i(x) \\ &= \sum_{i=1}^{N} f_i^{SE}(x) \cdot softmax(W_i f_i^{SE}(x)) \\ &= \sum_{i=1}^{N} f_i^{SE}(x) \cdot \frac{exp\left(W_i f_i^{SE}(x)\right)}{\sum_{i=1}^{N} exp(W_i f_i^{SE}(x))} \end{aligned} \quad (8)$$

where $f_{AE}(x)$ is the output of Attention-Based ensemble network, $f_i^{SE}(x)$ is the output of i-th base model, W_i is attention coefficient, N is the number of base models (N = 3 in this study). More detailed algorithm information can be found in Liu et al. (2022).

We used the site-scale ET estimation from three base models as input variables, and ground observed ET as the training target to train a neural network model with attention mechanism which was then used to estimate the global ET as a comparison.

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

3. Data and model validation

3.1 In situ observations

In this study, we take ground observed surface latent heat flux (LE) to calculate the ET values for training and validation. We used data from 129 flux tower sites (Table A1 and Fig. 2a.) in the FLUXNET2015 dataset (https://fluxnet.org/) (Pastorello et al., 2020) with the sampling frequency of half - hourly or hourly. These selected sites represent a wide range of major IGBP land cover types: cropland (CRO, 15 sites), deciduous broadleaf forests (DBF, 14 sites), evergreen broadleaf forests (EBF, 10 sites), evergreen needleleaf forests (ENF, 28 sites), grasslands (GRA, 28 sites), mixed forests (MF, 5 sites), open shrublands (OSH, 7 sites), savannas (SAV, 13 sites) and permanent wetlands (WL, 9 sites).

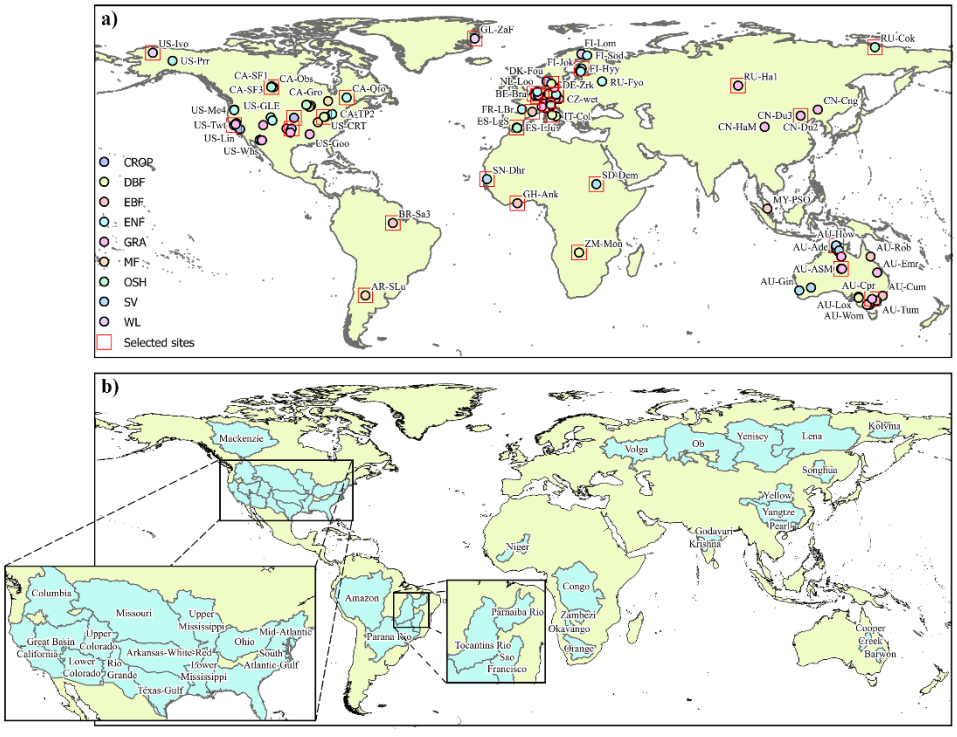

Figure 2. Locations of a) the 129 Flux sites and b) the 38 global catchments chosen for analysis in this study. The land cover types are identified based on the International Geosphere-Biosphere Programme (IGBP) biome classification. The red boxes indicate the locations of the 30 sites used in the independent validation.

To mitigate the effects of the uncertainty from data, we only chose the sites with high-quality data and rejected the non-observed data, missing values and data with energy closure less than 70% in the original dataset. Due to potential issues with eddy covariance technology and measurements under rainy conditions (Medlyn et al., 2011), we excluded rainy day samples to avoid errors. We also applied the Bowen ratio closure method to address the issue of energy imbalance in original observations (Foken, 2008; Twine et al., 2000). If the filtered data had more than 20% missing values in a month, the data

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

for that month was removed. The remaining data were then processed using linear interpolation and averaging to generate monthly-scale data.

3.2. Catchment- scale ET

At the catchment scale, the water balance method can be used to obtain a more reliable ET data (Pascolini-Campbell et al., 2020). The catchment-scale ET can be calculated as:

$$ET = P - Q - \frac{dS}{dt} \quad (9)$$

where P is precipitation, Q is runoff at the basin outlet, and dS/dt is the change in total water storage. We used the dataset of the water-balance-based evapotranspiration of global typical large river basins published by Ma et al. (2024a). This dataset spans a 34-year period from 1983 to 2016 and can be downloaded from the National Tibetan Plateau Data Center (Ma, 2024b). This data is derived from water balance methods combined with four different precipitation data sources (P), three types of terrestrial water storage change estimates (dS/dt) and observed flow data from control sites (Q). We selected 38 major catchments (>200,000 square kilometers) in this dataset as our validation data (Fig. 2b and Table A2).

3.3. Global- scale datasets

At the global scale, we collected 6 widely used ET products generated from different data sources, different forcing data, different calculation methods to evaluate model performance. (1) FLUXCOM (Jung et al., 2019; Tramontana et al., 2016) provides a latent heat dataset, with 0.0833° × 0.0833° resolution and time spanning a 75-year period from 1950 onwards generated, from energy flux measurements from FLUXNET eddy covariance towers and meteorological data utilizing only machine learning methods. (2) Process-based Land Surface Evapotranspiration/Heat Fluxes (PLSH) (Zhang et al., 2015) is a product with 0.0833° × 0.0833° resolution and time spanning a 32-year period from 1982 to 2013. The dataset is driven by satellite observations of photosynthetic canopy cover and surface meteorology inputs. (3) Global Land Evaporation Amsterdam Model (GLEAM) version 3.8a (Martens et al., 2017; Miralles et al., 2011) is an established ET product based on satellite and reanalysis data with 0.25◦ × 0.25◦ resolution and time spanning the 44-year period from 1980 to 2022. (4) GLEAM version 3.8b (Martens et al., 2017; Miralles et al., 2011) is also selected because of its different forcing data. It is based on only satellite data with 0.25◦ × 0.25◦ resolution and time spanning the 20-year period from 2003 to 2022. (5) Global Land Data Assimilation System (GLDAS) version 2.1 (Rodell et al., 2004; Beaudoing & Rodell, 2020), a product with 0.25° × 0.25° resolution and time spanning a 24-year period from 2001 onwards, is generated by combining data assimilation techniques with satellite and ground-based observations. (6) European Centre for Medium-range Weather Forecasts (ECMWF)-ERA5-Land product (Muñoz Sabater, 2019) is a reanalysis product based on multi-source data with 0.1◦ × 0.1◦ and time spanning a 75-year period from 1950 onwards.

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

In the input training data, variables P, Rs, SM, ST, Ta, VPD, WS are also sourced from the ERA5-land product. Since the ERA5-Land dataset only provides dew point temperature data, we used the following formula for calculation (Abbott & Tabony, 1985):

$$VPD = 0.6108\left[e^{\frac{17.27T}{T+237.3}} - e^{\frac{17.27T_d}{T_d+237.3}}\right] \quad (10)$$

Here, T and T_d represent temperature and dew point temperature respectively(℃). Although these data are also available in the in situ observations, we chose to use the global-scale dataset instead to avoid issues caused by inconsistencies in spatial scales between the training data and the calculation data (Xiao et al., 2008). We employed an inverse distance weighting method to extract data at the coordinates of the corresponding sites for training data.

Other variables LAI and NDVI are sourced from the Global Inventory Modeling and Mapping Studies (GIMMS) NDVI and LAI products (Cao et al., 2023; Li et al., 2023), which are half-month products with 0.0833° × 0.0833° resolution. We averaged the two values for the month to obtain the monthly average data. The IGBP variables are sourced from the MODIS Land Cover Climate Modeling Grid Product (MCD12C1) (Friedl & Sulla-Menashe, 2015), which is a spatially aggregated and reprojected version of the tiled MCD12Q1 product with 0.05° × 0.05° resolution. We resampled all these data to 0.1° spatial resolution, the same as the ERA5-Land product that contains the most input variables.

3.4. Model validation

3.4.1 Site and catchment scale validation

Given that both in situ observations and the total catchment ET calculated by the water balance method are relatively reliable datasets, we can directly compare the ET generated by our model with these datasets at site and catchment scale. At the site scale, we conducted k-fold cross-validation with k = 10. In each fold we selected 10% of the data at each site as the validation set and the remaining data as the training set. In addition, we carried out independent validation by selecting 30 independent sites excluded from the training data (Table A3 and Fig. 2a) to evaluate the spatial simulation performance of our Classifier-Guided Ensemble model. The selected independent validation sites cover multiple land cover types and most of the major regions. Also, we evaluated the ET estimation performance of our model on extreme samples sorted in ascending order within the 0th – 1st percentiles and 99th - 100th percentiles for six variables (LAI, NDVI, Rs, SM, Ta, VPD) to verify the extrapolation performance of the model. At the catchment scale, we used calculated catchment ET to compare with the total catchment ET estimated by our model and other global ET products.

To quantitatively validate the performance of our model, we used several statistical metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Coefficient of Determination (R^2), and Kling-Gupta Efficiency (KGE) (Table 1). MAE and RMSE are used to measure the closeness between the model results and observations. R^2 can be used to assess the correlations. KGE comprehensively considers the relative error, correlation coefficient and variance of the model results, providing a comprehensive assessment of the model's overall performance.

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

Table 1. The explanation of statistical metrics used to evaluate model performance in this study.

Statistic metrics	Unit	Equation	Suitable value
RMSE	mm month^{-1}	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$	0
MAE	mm month^{-1}	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$	0
R^2	NA	$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \mu_y)^2}$	1
KGE	NA	$KGE = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2}$ $\alpha = \frac{\mu_{\hat{y}}}{\mu_y}, \beta = \frac{\sigma_{\hat{y}}}{\sigma_y}$	1

Note. y_i and $\hat{y}_i$ represent observations and estimations, respectively, α is the bias ratio, β is the variability ratio, σ represents the standard deviation, μ represents the average.

3.4.2 Global scale validation

Taking into account the available years of each product, we analyzed the data for the period from 2003 to 2013, with the data of 2003 and 2004 serving as the training set and the data of remaining years as the validation set. At global scale, due to the lack of reliable ET observations, conventional validation methods are not feasible. Therefore, we used the three-cornered hat (TCH) method (Tavella & Premoli, 1994) to quantify the uncertainties of ET estimation. TCH is a reliable method for estimating the uncertainty of various time series products and has been used for validation of multiple datasets (Liu et al., 2021). We used the same formula as in (Xie et al., 2024). The time series of ET at each pixel can be expressed as:

$$X_k = X_t + \varepsilon_k, \forall k = 1,2,3,\dots,N \quad (11)$$

where k represents the index of the corresponding ET product, N is the total number of ET products included in the validation, X_t represents the true value and ε_k represents the error term.

The truth value of ET cannot be obtained, so we need to get the ε_k without X_t. To solve this problem, the TCH algorithm defines a product as a reference product to get a matrix $Y_{k,m}$ as:

$$Y_{k,m} = X_k - X_R = \varepsilon_k - \varepsilon_R, \forall k = 1,2,3,\dots,N-1 \quad (12)$$

where X_R is the arbitrarily selected reference product, $Y_{k,m}$ is an $M \times (N-1)$ matrix, M is the length of the time series of X. S, the covariance matrix of Y, is related to the unknown $M \times N$ covariance matrix of the individual noise Q as (Galindo & Palacio, 2003):

$$S = J \cdot Q \cdot J^T \quad (13)$$

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

$$J_{N-1,N} = \begin{bmatrix} 1 & 0 & \cdots & 0 & -1 \\ 0 & 1 & \cdots & 0 & -1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 \end{bmatrix} \quad (14)$$

Since the number of equations is still less than the number of unknowns, it is impossible to solve the equation. They used the Kuhn-Tucker theorem proposed by Galindo & Palacio (1999) to solve the con-strained minimization problem. The definition of the objective function and constraint function is:

$$F_{(r_{1N},\ldots r_{NN})} = -\frac{1}{K^2} \cdot \sum_{i<j}^{N} r_{ij}^2 \quad (15)$$

$$H_{(r_{1N},\ldots r_{NN})} = -\frac{|Q|}{|S| \cdot K} < 0 \quad (16)$$

where $r_{1N}, \ldots r_{NN}$ are the elements of the corresponding Q. They also used the initial conditions for the iteration proposed by Torcaso et al. (1998):

$$r_{iN}^0 = 0, i < N; \quad (17)$$

$$r_{NN}^0 = \frac{1}{2 \cdot S^*}, S^* = [1, \ldots, 1] \cdot S^{-1} \cdot (1, \ldots, 1)^T \quad (18)$$

Finally, the Q value is obtained by minimizing the objective function. More detailed information can be found in their paper (Galindo & Palacio, 1999, 2003; Torcaso et al., 1998; Xie et al., 2024).

3.4.3 ML model interpretability analysis

We used SHAP (SHapley Additive exPlanations) to analyze the interpretability of the ML classifier we used. SHAP was originally proposed in game theory (Shapley, 1952) and was used as a method to equitably distribute benefits by calculating each member's marginal contribution. Subsequently, it was used to compute the marginal contribution of each input variables to enhance the interpretability of machine learning (Lundberg & Lee, 2017).

The Python library of SHAP was used to calculate the SHAP values for each input feature. SHAP library is applicable for interpretability analysis of ensemble models that integrate multiple ML models. In this study, SHAP values were employed to enhance our understanding of how our ML classifier selects the 'dominant model'.

4. Results

4.1 Model evaluation with in situ observations

4.1.1 Model Evaluation using EC observations

To evaluate the model simulation performance, we conducted a ten-fold cross-validation at all selected EC sites (Fig. 3). First, we check whether the integrated base models could accurately estimate ET at site scale (Fig. 3b, c and e). It is found that both ML model and Hybrid model perform well with R^2 and KGE nearing 0.9, ensuring the reliability of ensemble

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

model while process-based algorithm exhibits relatively lower performance, with a KGE of 0.65, R^2 of 0.48, MAE of 17.41 mm month^{-1}, RMSE of 26.45 mm month^{-1}. Despite the lower accuracy of process-based algorithm at site scale, it is also integrated as the employed physically-founded equations enable process-based algorithm to maintain acceptable performance in data-sparse regions.

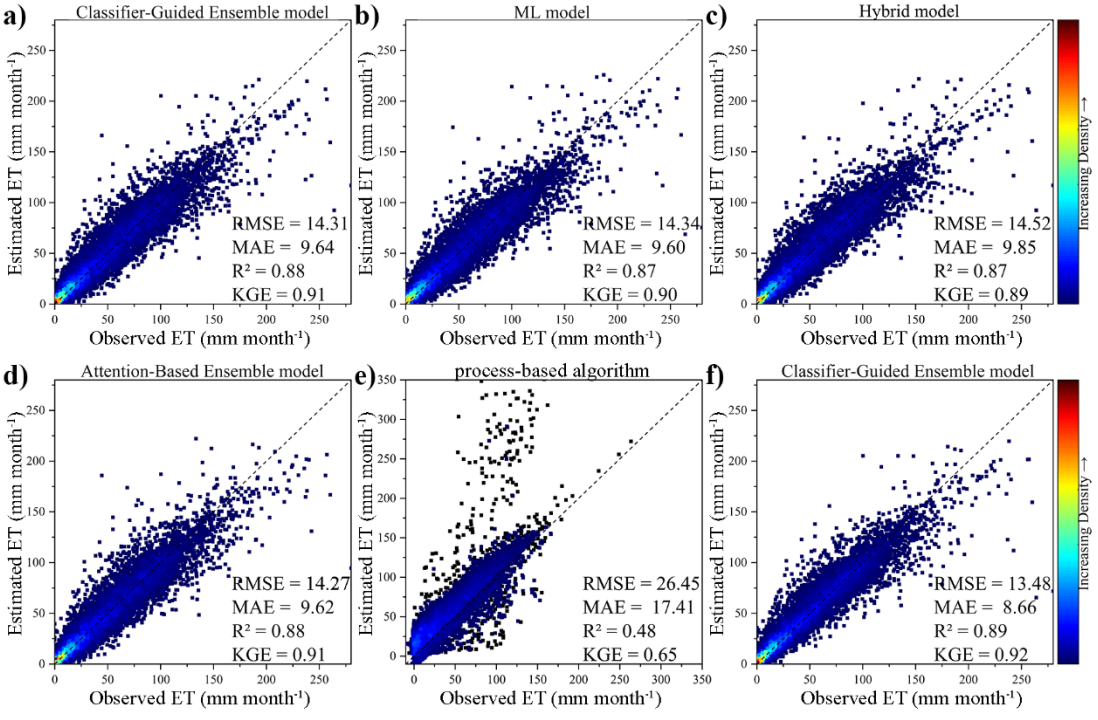

Figure 3. Performance of a) Classifier-Guided Ensemble model, b) ML model, c) Hybrid model, d) Attention-Based Ensemble model, e) process-based algorithm and f) Classifier-Guided Ensemble model (enhanced by additional global-scale training data) in ten-fold cross-validation.

The performance of our Classifier-Guided Ensemble model and the model with more training data are shown in Fig. 3a and 3f. Although when trained using only site-scale data, our Classifier-Guided Ensemble model performs well with a KGE of 0.91, R^2 of 0.88, MAE of 9.64 mm month^{-1}, RMSE of 14.31 mm month^{-1}, showing clear advantages over base models, it cannot significantly outperform the Attention-Based Ensemble model with the same KGE and R^2, lower RMSE of 14.27 mm month^{-1} and lower MAE of 9.62 mm month^{-1} (Fig. 3d). As illustrated in Fig. 3e, the process-based algorithm exhibits inferior performance to data-driven models at the site scale, resulting in its near exclusion from the classifier when trained only on in situ observations. Expanding the training dataset with global datasets enables the classifier to better recognize the scenarios where each base model performs best, with accuracy improving from 70% to 90%, particularly for process-based algorithm. Therefore, our model can achieve the best performance among all of the models used for comparison in the validation dataset, with a KGE of 0.92, R^2 of 0.89, MAE of 8.66 mm month^{-1}, RMSE of 13.48 mm month^{-1}.

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

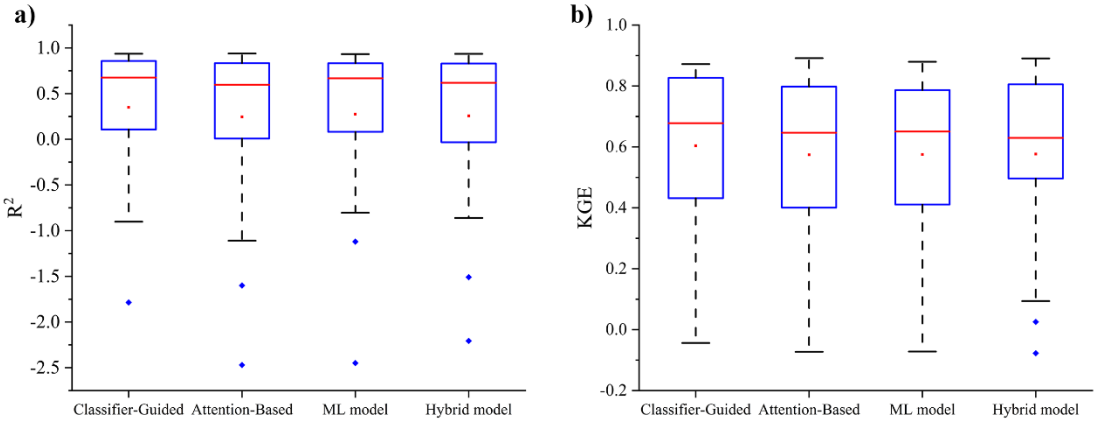

Figure 4. The a) R^2 and b) KGE of the Classifier-Guided Ensemble model, Attention-Based Ensemble model, ML model, and Hybrid model in independent validation. The red lines represent the median of the validation metrics, and the red dots represent the average values of the validation metrics.

The independent validation results shown in Fig. 4 and Table A4 also indicate that our model exhibits excellent generalizability in these independent sites. The ET estimation from Classifier-Guided Ensemble model achieves the best performance among all compared models, with a higher average R^2 of 0.35 and a higher average KGE of 0.60. The Classifier-Guided Ensemble model performs well in most of the selected independent sites, especially at the CH-Dav and US-ARM sites, as it can make better use of the process-based algorithm's extrapolation strengths, yielding better outcomes even when ML model and Hybrid model struggle, while the Attention-Based Ensemble model's results are closer to those of the ML model at most sites, resulting in poorer performance in independent validation.

The results from the k-fold cross-validation and independent validation indicate that our Classifier-Guided Ensemble model performs well in estimating ET at site scale, exhibiting better stability and generalizability, and the inclusion of global-scale data makes our model perform better.

4.1.2 Model Evaluation under different sites and vegetation cover conditions.

In situ observations were also used to evaluate the performance of models across different sites and land cover types, thereby validating the models' spatial simulation performance. The Taylor diagram is used to compare the performance of the models, where the two axes represent the root mean square error (RMSE, in mm $month^{-1}$), and the curves indicate the Pearson's correlation coefficient (r).

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

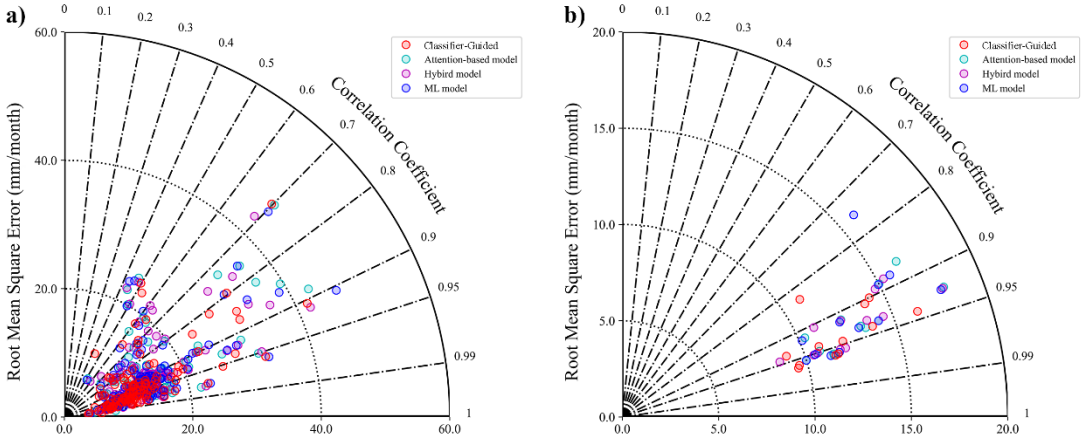

Figure 5. Taylor diagram that compares the performance of the four models with ground observations for a) different sites and b) different land cover types.

As Fig. 5a and Table 2 demonstrates, Classifier-Guided Ensemble model performs the best among the four models in the majority of the site, with lower average RMSE of 14.55 mm month^{-1} (Attention-Based Ensemble model: 15.41 mm month^{-1}, Hybrid model: 15.52 mm month^{-1}, ML model: 15.46 mm month^{-1}), and higher average correlation coefficient(r) of 0.90 (Attention-Based Ensemble model: 0.88, Hybrid model: 0.88, ML model: 0.87). Classifier-Guided Ensemble model also demonstrates greater performance than the other three models in the majority of land cover types (Fig. 5b and Table 3), with lower average RMSE of 16.88 mm month^{-1} (Attention-Based Ensemble model: 17.40 mm month^{-1}, Hybrid model: 17.38 mm month^{-1}, ML model: 17.94 mm month^{-1}), and higher average correlation coefficient(r) of 0.93 (Attention-Based Ensemble model: 0.92, Hybrid model: 0.93, ML model: 0.92). Classifier-Guided Ensemble model outperforms other models across most land cover types, with the exception of the CSH, where its performance is slightly inferior to Attention-Based Ensemble model and Hybrid model.

Table 2. Performance of different ET models as indicated by averaged RMSE and R^2 for all sites.

Model	RMSE (mm month^{-1})	R^2
Classifier-Guided Ensemble model	14.55	0.90
Attention-Based Ensemble model	15.41	0.88
Hybrid model	15.52	0.88
ML model	15.46	0.87

Further, we notice that the Attention-Based Ensemble model assigns more ‘attention’ to the ML model, since the ML model performs best at the site scale among the three base models. Therefore, the performance of the Attention-Based Ensemble model is inferior to that of the hybrid model under different land cover types, while Classifier-Guided Ensemble model can fully utilize the characteristics of the base models and get better results, which is consistent with the conclusion

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

from independent validation. In summary, the proposed model better fits the data from different sites and different land cover types, which demonstrates the effectiveness of our ensemble method.

Table 3. Performance of different ET models as indicated by RMSE and R^2 for all IGBP land cover types.

IGBP	Classifier-Guided Ensemble model		Attention-Based Ensemble model		Hybrid model		ML model	
	RMSE (mm month^{-1})	R^2	RMSE (mm month^{-1})	R^2	RMSE (mm month^{-1})	R^2	RMSE (mm month^{-1})	R^2
ENF	9.57	0.96	9.99	0.96	10.44	0.95	9.99	0.96
EBF	14.23	0.90	15.01	0.89	14.70	0.89	14.96	0.89
DBF	11.68	0.96	11.45	0.96	12.13	0.96	11.76	0.96
MF	9.49	0.96	10.79	0.95	10.65	0.95	10.52	0.95
CSH	56.82	0.84	54.77	0.83	52.81	0.88	60.26	0.80
OSH	9.08	0.94	10.33	0.92	10.97	0.91	10.14	0.92
SV	12.80	0.95	12.99	0.95	13.33	0.94	12.92	0.95
GRA	12.30	0.94	13.62	0.93	13.82	0.93	13.45	0.93
CROP	15.92	0.94	17.62	0.93	17.54	0.93	17.44	0.93
Average	16.88	0.93	17.40	0.92	17.38	0.93	17.94	0.92

4.1.3 Model Evaluation using extreme samples

In order to verify the extrapolation performance of the models for extreme samples, we compared the performance of these models for multiple extreme samples. The heatmaps in Fig. 6 indicate that for the majority of these extreme samples, our Classifier-Guided Ensemble model can accurately estimate ET.

For the 99th – 100th percentiles of the VPD in particular, Classifier-Guided Ensemble model performs significantly better than the other models, with a KGE of 0.66, R^2 of 0.36, while the KGE of other models is below 0.3 and the R^2 is less than 0. This shows that Classifier-Guided Ensemble model has the potential to efficiently select the 'dominant model' to achieve good results even when these existing ML models perform poorly. Under the cases of high Ta, high Rs and low Ta, Classifier-Guided Ensemble model performs significantly better than the other models. Only under the three cases of low VPD, low Rs and low LAI, Classifier-Guided Ensemble model is not as good as the other models and the difference is not significant, with the KGE being 0.08, 0.02 and 0.08 lower than that of the Attention-Based Ensemble model, respectively. In most extreme cases, Classifier-Guided Ensemble model can yield stronger extrapolation performance than individual base models or other ensemble models, providing more accurate ET estimates under extreme weather events.

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

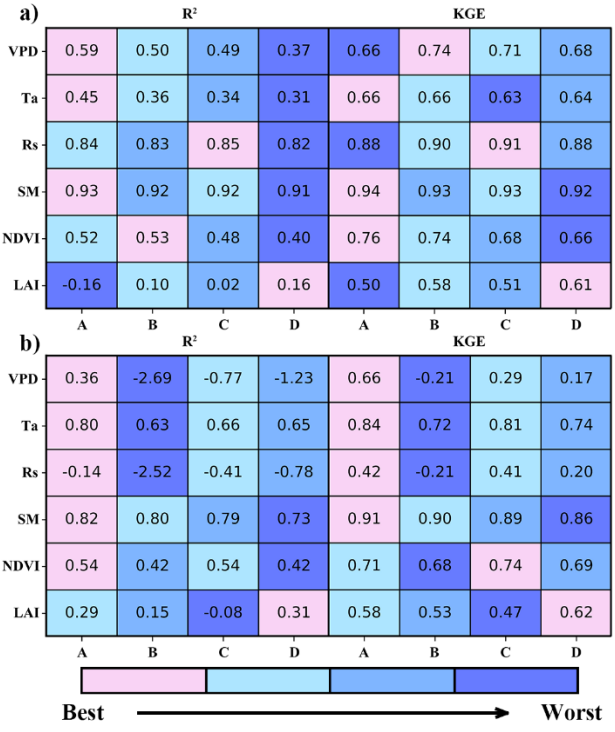

Figure 6. The comparison of different models (A Classifier-Guided Ensemble model, B Attention-Based Ensemble model, C ML model, D Hybrid model) under extreme conditions in the form of heatmaps. a) and b) represent the extreme samples sorted in ascending order within the 0th – 1st percentiles and 99th - 100th percentiles, respectively.

4.2 Model evaluation at catchment scale

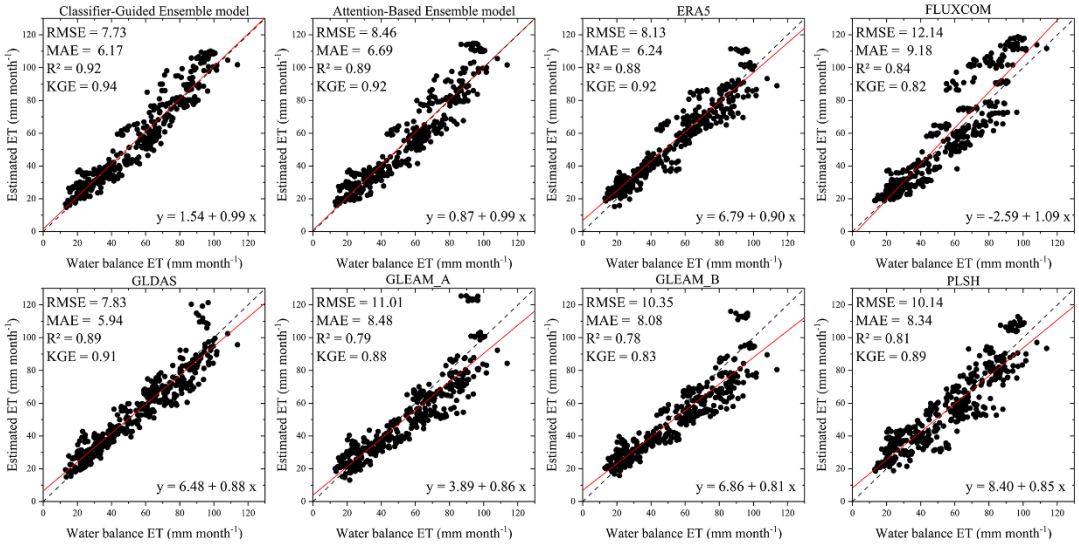

Figure 7. Scatterplot for the relationship between estimated ET and water balance ET (each point represents a catchment over a one-year period).

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

To validate the model performance for the catchment-scale application, we used six ET products, Classifier-Guided Ensemble model, and Attention-Based Ensemble model to estimate ET in each catchment and compare them to the water balance ET dataset. The results show that Classifier-Guided Ensemble model performs better than other comparison models and products (Fig. 7), with a KGE of 0.94, R^2 of 0.92, MAE of 6.17 mm month^{-1}, RMSE of 7.73 mm month^{-1} and the slope of the regression of estimated ET versus water balance ET for our model (0.99) is closer to 1 than that of the other ET products and models. The Attention-Based Ensemble model also performs better than most of the ET products used, with a KGE of 0.89, but the results are not as good as those of our model, ERA5-Land, and GLDAS ET products as it exhibits higher RMSE and MAE. This proves that our model is in better agreement with the catchment water balance ET.

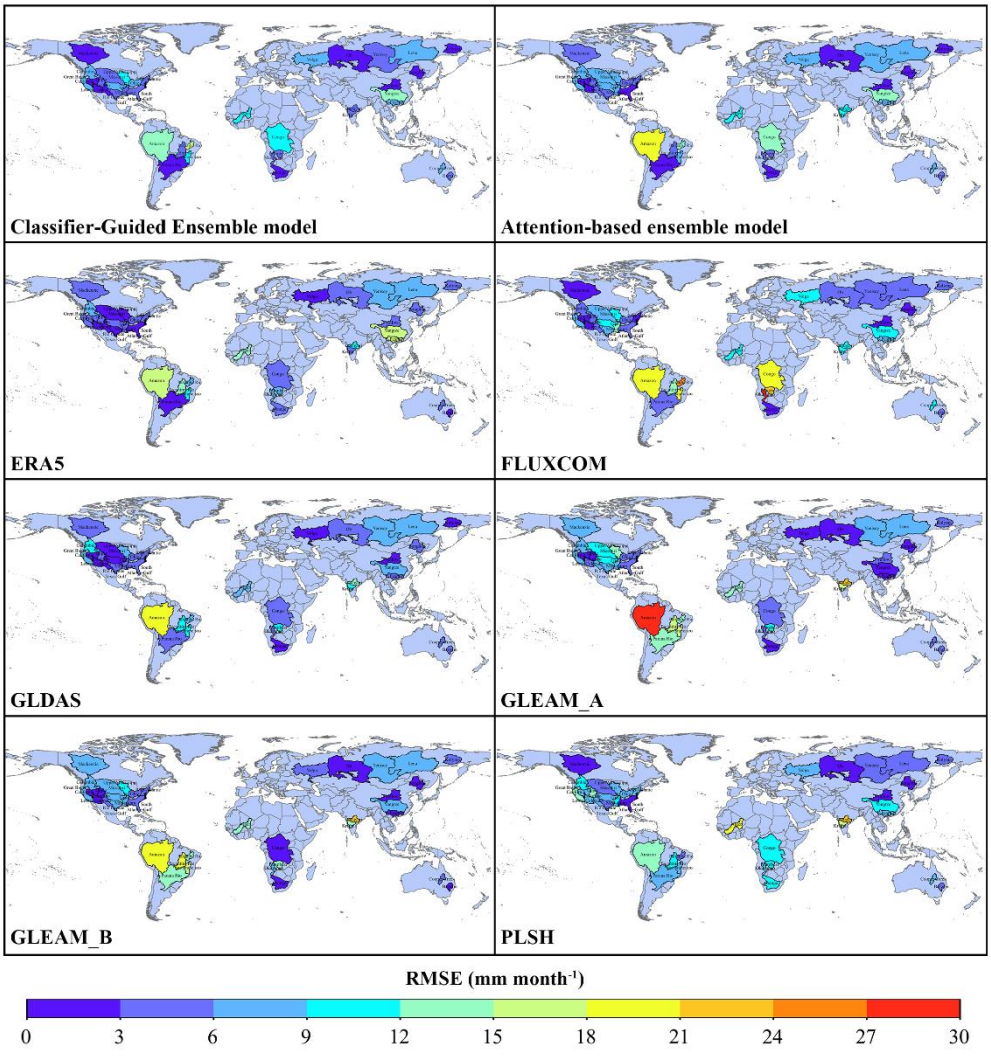

Figure 8. Distributions of the root-mean-square error (RMSE) for the 38 catchments.

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

The spatial distribution of RMSE across 38 catchments (Fig. 8) further demonstrate the superior performance of our model. Compared to the Attention-Based Ensemble model, our model shows significant improvement in RMSE for catchment 'Mackenzie', 'Rio Grande', 'Lower Colorado', 'Amazon', 'Kolyma ', 'Godavari', 'Krishna', 'Orange', 'Cooper Creek' and 'Barwon', with a reduction in RMSE of over 20%. Especially in catchment 'Godavari', the RMSE of our model is 5.59 mm month^{-1} and the RMSE of Attention-Based Ensemble model is 12.02 mm month^{-1}, with a difference of more than 50%. In some other catchments ('Mid-Atlantic', 'Ohio', 'Upper Colorado', 'Ob' and 'Yellow'), the RMSE of the Attention-Based Ensemble model is over 20% lower than that of our model. Compared to another well-performing GLDAS product, there are large differences between our model and GLDAS due to the difference in calculation methods, with each of these two methods having lower RMSE in 19 catchments. In summary, while there are some differences in ET estimation among our model, the Attention-Based Ensemble model, and other products in different catchments, our model is in better agreement with the catchment water balance ET in the majority of catchments.

4.3. Model evaluation at global scale

To further validate the performance of our model on larger spatial scales, we compared the multi-year average ET estimated by Classifier-Guided Ensemble model and Attention-Based Ensemble model, and other ET products. We primarily compared our Classifier-Guided Ensemble model with Attention-Based Ensemble model to validate the performance of our ensemble method and with FLUXCOM, which is an ET product generated based on pure ML models, to evaluate the differences between ensemble model and pure ML models. Additionally, other ET products were included as reference for the analysis. Some ET products, like FLUXCOM, contain missing values in certain regions, and these regions were excluded from the comparison.

4.3.1 Evaluation of multi-year average ET estimates

Fig. 9 shows the spatial distribution of the multiyear (2005-2013) mean global ET estimates for the ET models and products. Our Classifier-Guided Ensemble model shows expected global patterns of ET and all of these ET models and products generally show similar spatial pattern. ET values are relatively higher in mid-latitude regions near the equator, including the Amazon Basin, the Congo Basin, and Southeast Asia, while lower ET values are shown in some arid regions, including the Sahara Desert and Central Asia, as well as in high-latitude alpine regions, including northern Russia, and northern Canada. The multiyear mean ET ranges from 46.99 mm month^{-1} to 49.69 mm month^{-1}, with FLUXCOM having the highest average ET and GLEAM_B having the lowest average ET.

Despite their great consistency in spatial patterns, there are still some regional discrepancies detected among these datasets. Compared with other products, the Attention-Based Ensemble model shows the largest discrepancy in tropical regions, as it provides lower ET estimates, with almost no regions exhibiting ET exceeding 120 mm month^{-1}. This may be attributed to the ML model's limited ability to estimate extreme high values, as well as the potential underestimation of ET by the ensemble model (Cai et al., 2024). Compared with the Attention-Based Ensemble model, our Classifier-Guided

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

Ensemble model shows better spatial consistency with other ET products and avoids the underestimation of high values. The latitudinal average ET distribution for each dataset also confirms this conclusion. The Attention-Based Ensemble model shows lower ET estimates in both high-ET and low-ET regions and the FLUXCOM product tends to overestimate ET in low-latitude regions, particularly in areas slightly below the peak values, while the ET profile estimated by our model shows
435 improvement in these areas. Overall, our model generally performs well and provides a reasonable global ET estimate, indicating a distinct improvement in generalizability with the introduction of ML classifier.

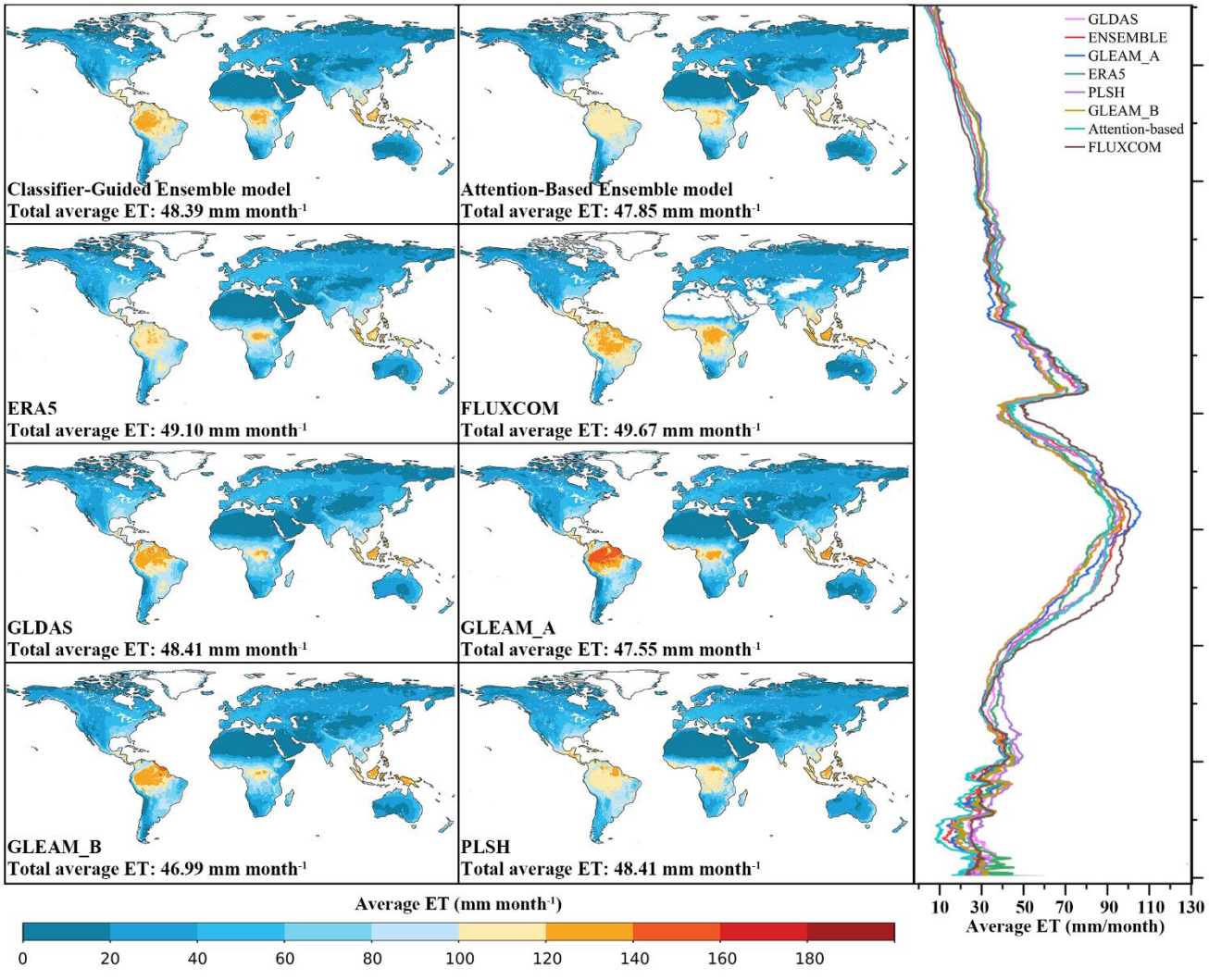

Figure 9. Average annual land evapotranspiration from 2005 to 2013 for Classifier-Guided Ensemble model, Attention-Based Ensemble model and other ET products. The latitudinal profiles of these datasets are shown in the right panel.

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

4.3.2 Uncertainty analysis at global scale

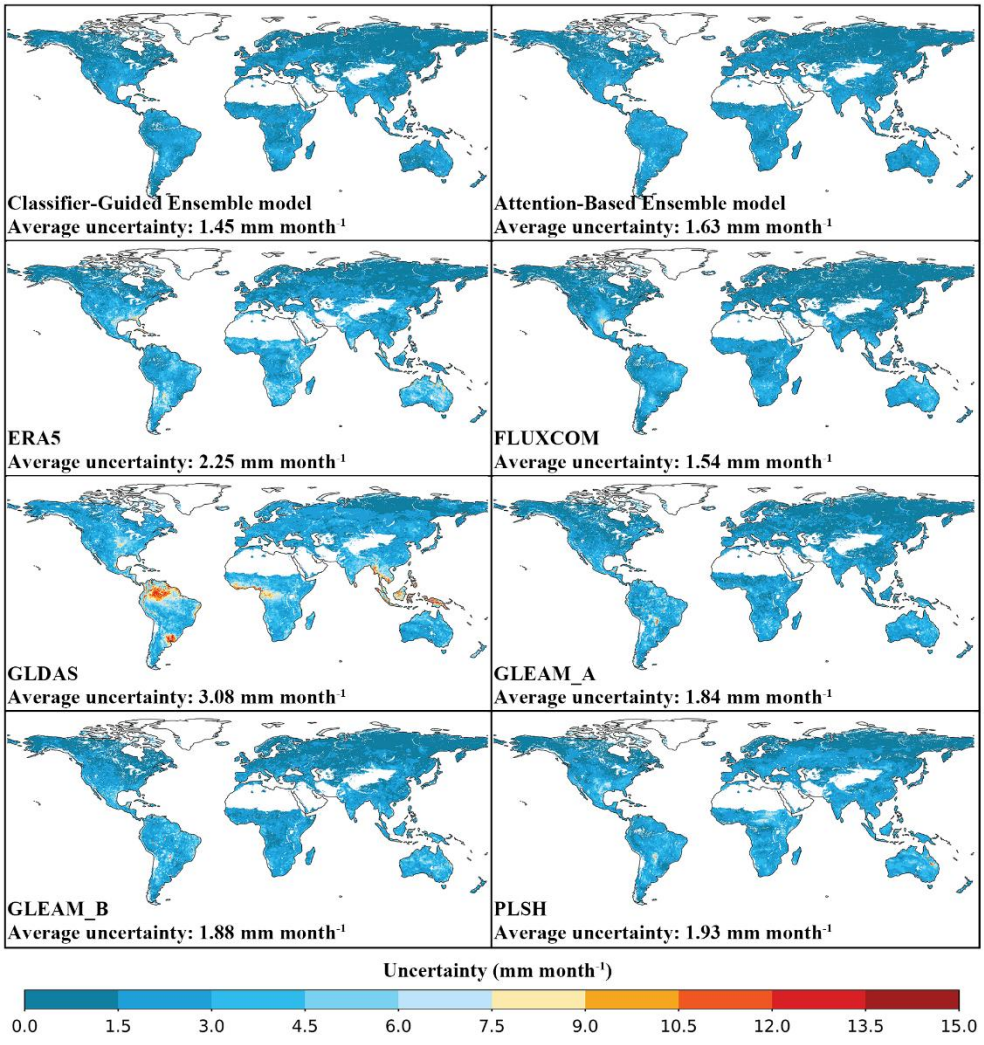

Figure 10. Average monthly uncertainty from 2005 to 2013 for Classifier-Guided Ensemble model, Attention-Based Ensemble model and other ET products.

The proposed model also demonstrates strong stability in the uncertainty analysis. Fig. 10 illustrates the uncertainty distribution estimated from the TCH method for eight global ET estimates during the period from 2005 to 2013. All of these datasets exhibit high stability, with less than 5 mm month^{-1} mean uncertainty in most areas. The lowest mean uncertainty is achieved by Classifier-Guided Ensemble model (1.45), followed by FLUXCOM (1.54), Attention-Based Ensemble model (1.63), GLEAM_A (1.84), GLEAM_B (1.88), PLSH (1.93), ERA5 (2.25), GLDAS (3.08). The uncertainty distribution of these datasets also shows a similar spatial pattern and typically high uncertainty is found in low-latitude areas. Many studies have investigated the uncertainty of these ET products. Xie et al. (2024) used the TCH method to assess the uncertainty of

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

several products from 2003 to 2015 and found that FLUXCOM had the lowest uncertainty. Zhu et al. (2022) and Li et al., (2022) also evaluated several products and found that the GLEAM product exhibited lower uncertainty. The uncertainty of these products we calculated is consistent with the results of these previous studies.

Compared to Attention-Based Ensemble model, our model shows lower uncertainty in high ET regions near the equator and shows lower uncertainty compared to FLUXCOM in the southern North America and Australia. These results confirm preceding analysis that our Classifier-Guided Ensemble model consistently perform well at global scale, showcasing the potential of our ensemble method to enhance the generalizability.

5. Discussion

In this work, by developing a novel ML Classifier-Guided Ensemble ET model, we provide a simple but effective way to integrate different base models (process-based algorithm, ML-based ET model, and hybrid model) to estimate ET or similar variables lacking reliable global observations. Through the introduction of ML classifier, Classifier-Guided Ensemble model can utilize the distinct advantages of the three base models with better performance at multiple spatial scales. Compared with individual base models and Attention-Based Ensemble model, Classifier-Guided Ensemble model fit ET observations better, especially in extreme samples and under different sites and vegetation cover conditions, demonstrating improved generalizability and avoiding the underestimation of high values compared to traditional ensemble models. At catchment scale, ET estimates from Classifier-Guided Ensemble model show a greater agreement with catchment ET calculated from water balance, with performance comparable to other widely used ET products (ERA5, FLUXCOM, GLDAS, GLEAM_A, GLEAM_B, and PLSH). At global scale, the evaluation of multi-year average ET estimates and uncertainty analysis indicate that Classifier-Guided Ensemble model can provide a reasonable and stable global ET estimate. The main advantage of Classifier-Guided Ensemble model is the improved generalizability, which is primarily attributed to the introduction of ML classifier due to the ML Classifier's capacity to include a broader range of training data and to select the appropriate model at each pixel, especially for process-based algorithm.

5.1 Evaluation of the Effectiveness of ML classifier.

As the additionally incorporated ML classifier is the core of the proposed ensemble framework, we further validated its effectiveness at both site and global scales. Fig. 11a shows that in 'process-based algorithm-dominated' type derived from in situ observations, our Classifier-Guided Ensemble model achieves better results than other models, with a KGE of 0.94, R^2 of 0.94, MAE of 5.07 mm month-1, RMSE of 8.71 mm month^{-1}. As demonstrated by the results in Section 4.1.1, the process-based algorithm has a poor accuracy of ET estimation at site scale, with a KGE of 0.65, so ensemble model based on in situ observations cannot take good advantage of the process-based algorithm. In Attention-Based Ensemble model, the process-based algorithm contributes much less than the other two models, so neither ML model nor Attention-Based

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

Ensemble model can outperform our Classifier-Guided Ensemble model when the ET estimated by process-based algorithm is the closest to the observed ET.

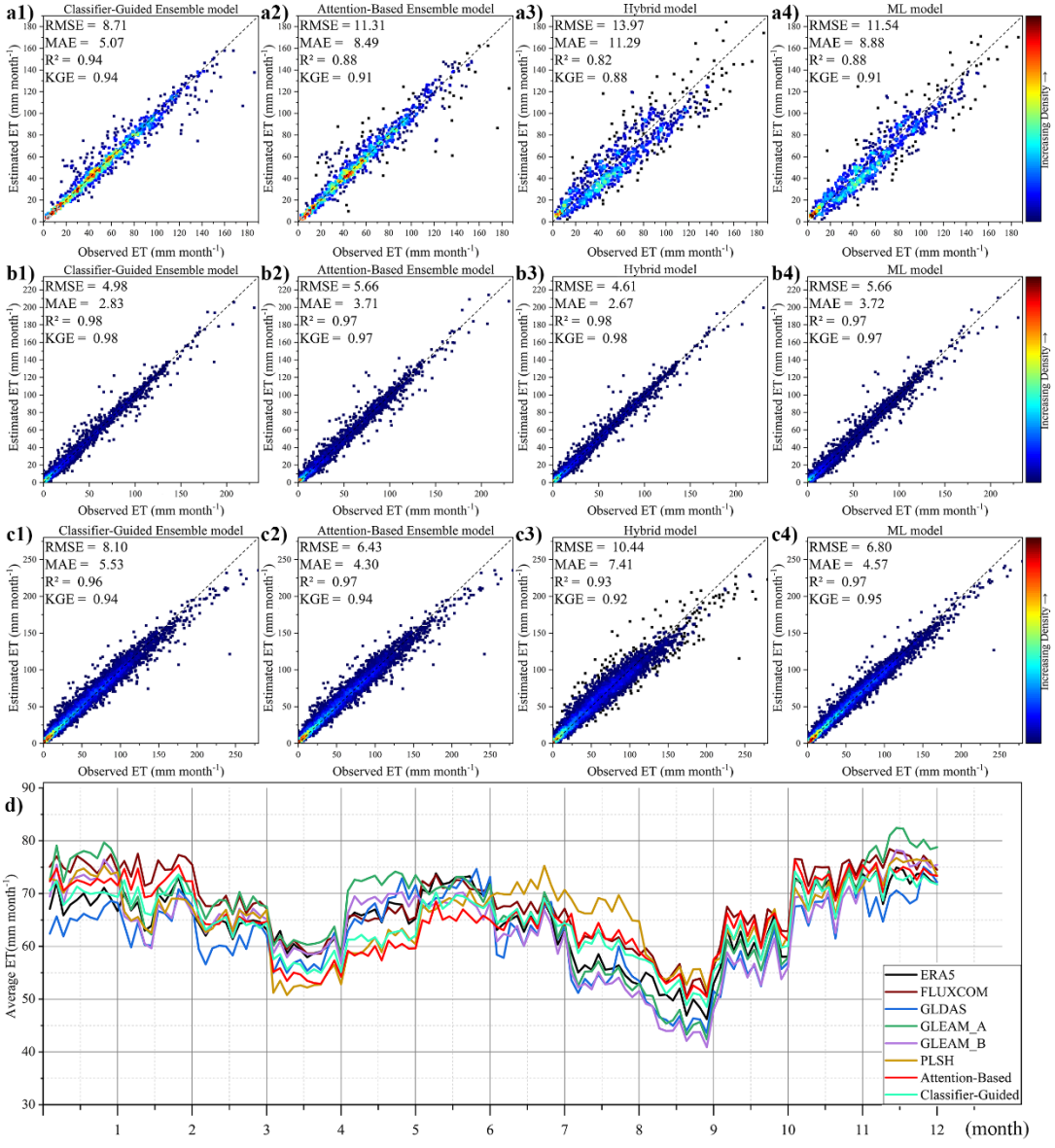

Figure 11. The comparison of model performance among 1) Ensemble model, 2) Attention-Based Ensemble model, 3) Hybrid model and 4) ML model under a) 'process-based algorithm-dominated' type, b) 'Hybrid model-dominated' type, c) 'ML model-dominated' type based on in situ observations. d) The comparison of global average monthly land evapotranspiration from 2005 to 2013 under 'process-based algorithm-dominated' type.

We also analyzed the case of two other types: 'hybrid model-dominated' and 'ML model-dominated' derived from in situ observations. In 'hybrid model-dominated' type (Fig. 11b), the hybrid model performs the best as expected, with a KGE

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

of 0.98, R^2 of 0.98, MAE of 2.66 mm $month^{-1}$, RMSE of 4.61 mm $month^{-1}$. Classifier-Guided Ensemble model also performs better in this case compared to Attention-Based Ensemble models, with a KGE of 0.98, R^2 of 0.98, MAE of 2.83 mm $month^{-1}$, RMSE of 4.92 mm $month^{-1}$. The Attention-Based Ensemble model pays more 'attention' to the ML model which has higher accuracy at site scale, so in 'hybrid model-dominated' type, its results are closer to the ML model, while our model can use hybrid model in most points based on the results of the ML classifier, leading to better results for our models. However, in 'ML model-dominated' type (Fig. 11c), Attention-Based Ensemble model performs better than Classifier-Guided Ensemble model, with a higher R^2 of 0.97, lower RMSE of 6.43 mm $month^{-1}$ and MAE of 4.30 mm $month^{-1}$. Classifier-Guided Ensemble model cannot perform as well as the ML model because the accuracy of ML classifier is not 100%, while the Attention-Based Ensemble model gets better performance by combining the results of the three models.

Therefore, although our model performs well in estimating ET at various scales, there are still some limitations. The core of our model is to select the potential optimal model as 'dominant model' for each pixel as determined by the ML classifier, so in regions where a single model already achieves the best results, Classifier-Guided Ensemble model does not improve performance. In this case, other ensemble model, such as Attention-Based Ensemble model, performs better, as they can improve performance by integrating multiple models.

At global scale, we also conducted additional analysis and validation of the results of the ML classifier within our model. Since ML model and Hybrid model perform well at global scale, while process-based algorithm has lower overall accuracy, we focused on whether the ML classifier can identify pixels where the process-based algorithm performs well and use it to improve the estimation robustness in these areas. Fig. 11d shows the line chart of the monthly mean ET series for all datasets at the points corresponding to 'process-based algorithm-dominated' type derived from the ML classifier. It is found that the Attention-Based model yields the lowest ET estimation among all datasets around May and FLUXCOM overestimates ET around February, September, October and November, while our Classifier-Guided Ensemble model shows improvements in these cases. In comparison to the Attention-Based Ensemble model's performance (R^2 with the various products as follows, ERA5: 0.60, FLUXCOM: 0.71, GLDAS: 0.29, GLEAM_A: 0.54, GLEAM_B: 0.51, and PLSH: 0.73), our model's results are closer to these widely used ET products (R^2 with the various products as follows, ERA5: 0.85, FLUXCOM: 0.72, GLDAS: 0.60, GLEAM_A: 0.67, GLEAM_B: 0.71, and PLSH: 0.84). This also demonstrates that using the process-based algorithm instead of ML models in these regions has led to an improvement in the reliability of the ET estimation. Overall, the introduction of the ML classifier did improve the performance of our model at both site and global scale.

5.2 Interpretability of machine learning used in Classifier-Guided Ensemble model

For machine learning-based models, improvement of the model performance is important, but the interpretability of the model is equally crucial. Especially for our model, the explanation of ML classifier can give us more insight into how it selects the 'dominant model', and under which meteorological conditions a particular model is preferred to be selected at both global and site scales, providing valuable support for future research. Interpretable machine learning models are gaining

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

increasing attention, and various methods, such as LIME and SHAP, have been widely used to explain various machine learning models (e.g. Chakraborty et al., 2021; Chu et al., 2024; Eskandari et al., 2024). In this study, we used SHAP values to analyze the interpretability of the ML classifier within Classifier-Guided Ensemble model and we used both site-scale and global-scale data to calculate SHAP values.

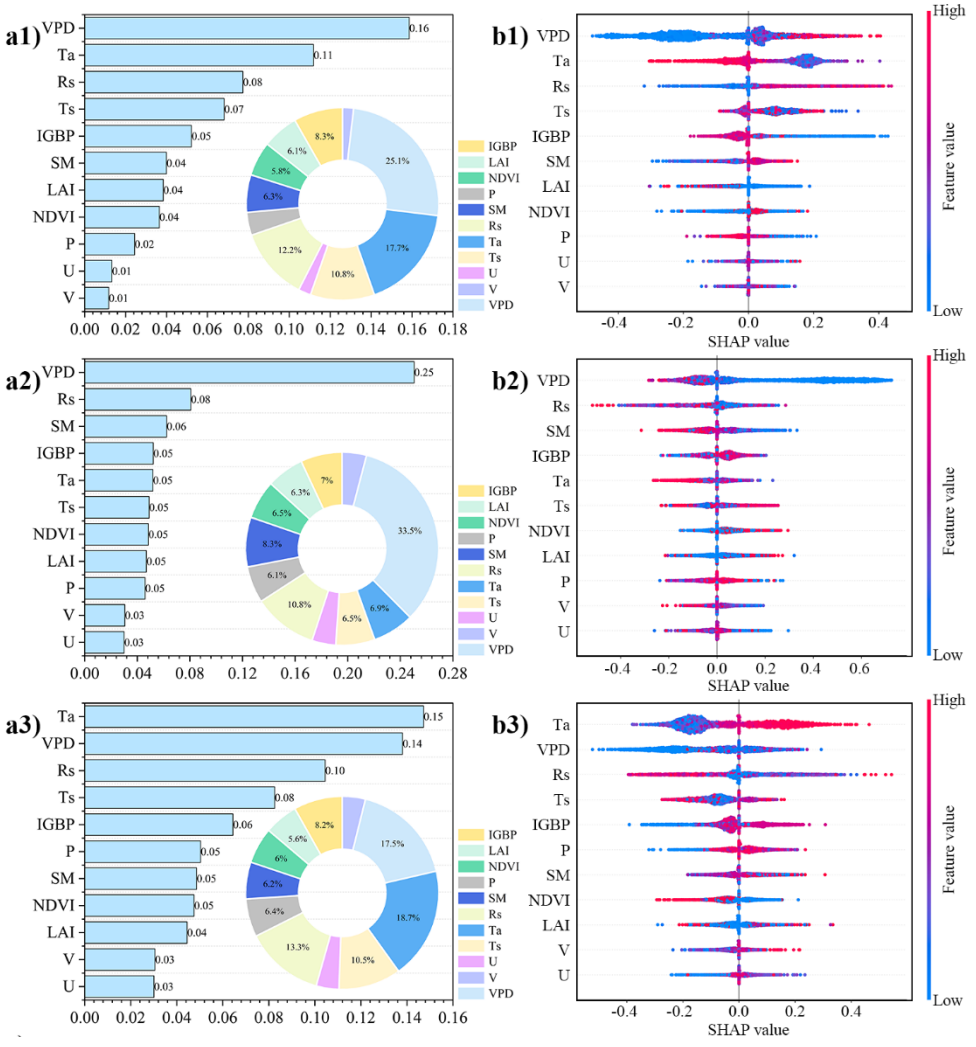

Figure 12. a) The bar plot and b) summary plot of the SHAP values for 1) 'process-based algorithm-dominated' type, 2) 'Hybrid model-dominated' type and 3) 'ML model-dominated' type. The bar plot exhibits the mean absolute SHAP values for each covariate and the summary plot exhibits the distribution of SHAP values.

The contribution of different covariates to the results varies across the different classes (Fig. 12a). As Fig. 13a shows, the covariates VPD, Ta, and Rs have a higher contribution, while NDVI, LAI, P, U, and V have a lower contribution. For 'process-based algorithm-dominated' type, the contribution of VPD accounts for 25.09%, followed by Ta (17.68%), Rs (12.22%), Ts (10.78%), IGBP (8.25%), SM (6.31%), LAI (6.08%), NDVI (5.77%), P (3.86%), U (2.10%), V (1.85%). For

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

'ML model-dominated' type, the distribution of SHAP values shows some similarity to 'process-based algorithm-dominated' type (Fig. 13b), with Ta having the highest contribution (18.67%) and VPD the second highest (17.52%). Previous studies have demonstrated that variables VPD, Ta, and Rs make significant contributions to the estimation of ET, whether using machine learning methods or the process-based algorithm (Mu et al., 2011; Shang et al., 2023). The similarity in the contribution distribution of ML model and process-based algorithm may be attributed to the fact that they directly estimate ET, while differences may result from their different algorithms.

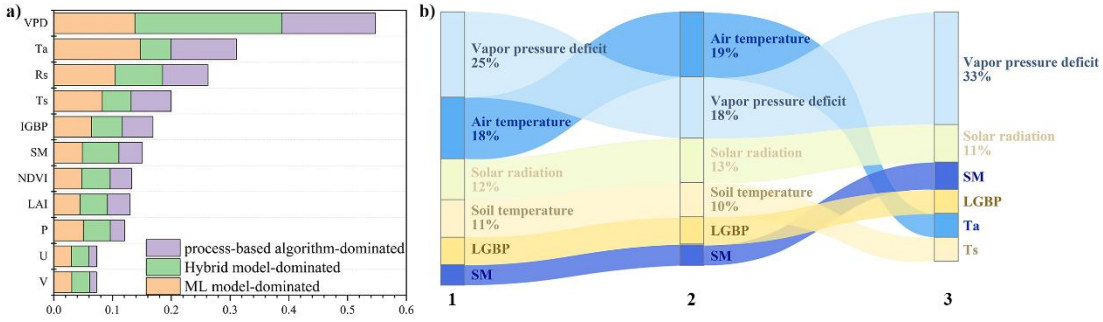

Figure 13. a) The bar plot of the mean absolute SHAP values for three types and b) the changes in variable contributions across 1) 'process-based algorithm-dominated' type, 2) 'ML model-dominated' type, 3) 'Hybrid model-dominated' type.

For 'Hybrid model-dominated' type, it is still VPD that dominates the feature contribution, with its proportion rising to 33.39%. The distribution of covariates under 'Hybrid model-dominated' type is somewhat distinct from 'process-based algorithm-dominated' type and 'ML model-dominated' type, with a higher contribution from SM and a slightly reduced contribution from Ta (Fig. 13b). Machine learning was used in the hybrid model for the estimation of r_s, and it was strongly correlated with water content (VPD, SM), and temperature (Ta) (Gan et al., 2018; Leuning et al., 2008; Mallick et al., 2015). In the hybrid model ML-GS developed by Shang et al., (2023), the distribution of SHAP values for their input variables is similar to that in our 'Hybrid model-dominated' type. Therefore, it can be seen that the variables with a higher contribution when selecting the 'dominant model' in the ML classifier have a certain correlation with those having a higher contribution when estimating ET.

The Fig. 12b shows the summary plot, where the horizontal axis represents the SHAP values, indicating the contribution of each covariate and color represents the magnitude of the variable values, with redder colors indicating larger variable values. The summary plot illustrates that, in certain conditions, a specific base model may have higher SHAP values, indicating its greater likelihood of being selected. For instance, under high VPD, low Ta and high Rs conditions, process-based algorithm is more likely to be selected. Also, there is a tendency to select Hybrid model under conditions of low VPD, low SM and under high Ta conditions, the ML model tends to be selected. The results in the summary plot show some correlation to the extreme samples analysis (Section 4.1.3) that for the 0th – 1st percentiles of the VPD and SM, hybrid model outperforms ML model, making hybrid model the preferred choice. Despite these clear conditions, model selection

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

remains unclear in some cases, particularly for ML model and Hybrid model, possibly due to their inter-correlation as both of them are based on ML method.

In summary, the interpretability analysis provides insights into the covariates with the high contribution to model selection, as well as scenarios where the three base models are more likely to be selected. However, the model selection for some cases such as high VPD, low radiation, etc. is not fully determined and further research is needed to explore the deeper mechanisms of model selection.

5.3 Uncertainties of Classifier-Guided Ensemble model

Despite better performance than other models in estimating ET at multiple spatial scales, there are still uncertainties in our model. First, the inclusion of a ML classifier in our ensemble model may introduce uncertainties. Although we used multiple global ET products as references when adding global-scale training data, these are not as reliable as ET observations, so we are still unsure of the accuracy of classification. Moreover, the classification results of machine learning model are not completely accurate, leading to the fact that the ML classifier does not guarantee optimal model selection. To minimize the impact of this issue, we have chosen models that have been shown to be excellent for global ET estimation in previous studies, so even if the ML classifier does not accurately choose the optimal model as 'dominant model' at some pixels, the results of the other models will not differ significantly.

Second, the base models selected may introduce uncertainties. We chose to integrate ML-based ET models, process-based algorithms, and hybrid models. Although process-based algorithms have been widely used for ET estimation, there is still no widely recognized optimal method for the parameterization of some ET processes (Jiménez et al., 2018; Mu et al., 2011). While ML-based ET models perform well in regions with sufficient data, they tend to have poor generalizability in regions with limited data and may suffer from local optima or overfitting problems (Koppa et al., 2022; Yuan et al., 2020). For hybrid model, there are also uncertainties in the synergy between physical laws and machine learning (Shang et al., 2023).

Lastly, our model contains many input variables that incorporate multiple data sources including: satellite data, reanalysis data, and in situ observations, which may introduce uncertainty. At site scale, water flux observations obtained by the eddy covariance method have inherent random errors from the measuring instruments (Mizoguchi et al., 2009). According to the previous study, the biases between the reanalysis data and the satellite meteorological data (Rienecker et al., 2011) and the inconsistency between in situ observations and global-scale data also introduces uncertainty (Cao et al., 2021). We have made efforts to reduce uncertainties caused by data inconsistencies by replacing in situ covariate observations with the corresponding data from global-scale datasets at the same coordinates, but the uncertainties remain unavoidable.

5.4 Future perspectives

Despite the uncertainties and limitations, our model offers a new perspective on integrating multiple models and utilizing the complementarity between ML models and physical models, and there remains potential for further improvement. First,

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

various alternatives still exist for the selection of base models. For example, there are many other ML models that can be chose and have their own advantages, such as some deep learning models (LSTM, CNN, etc.). These deep learning models or their improved forms have been shown to effectively estimate ET in previous studies (Guo et al., 2024; Karbasi et al., 2022; de Oliveira e Lucas et al., 2020). There are also some ET estimation methods based on other frameworks without using the three base models we chose, such as a Bayesian-driven ensemble learning method (Ochege et al., 2024). Better results might be obtained by choosing different models or by integrating more models, which could be a potential direction for future research.

Second, more rigorous and reasonable methods are needed for model integration. One of the advantages of the proposed framework is that we can add global data as training data even when there are no global-scale observations, which offers a novel approach to improving the generalizability of model, but how to identify the 'dominant model' from additional data remains a problem. If there is a more rigorous and reasonable method to optimize the model selection, it can not only improve the reliability of the final ET estimation, but also contribute to the study of the regions where each base model demonstrates its advantages. Additionally, after the ML classifier has selected the 'dominant model' at each pixel, if we discard the simple use of the selected model and adopt more advanced methods for model integration, the results of Classifier-Guided Ensemble model may be further improved. We have tried optimizing the results under three different classes using both genetic algorithms and machine learning models, but the results were not as good as expected. Since these methods are still limited by in situ observations, they do not perform as well as using the original models directly when upscaled to a global scale. It is still worth exploring how to better utilize these models.

Finally, more effective integration of machine learning and physical models has the potential to further improve ET estimation. Since physical models have higher interpretability and extrapolation capabilities, while machine learning can utilize data more effectively, how to better integrate the complementary advantages of them is a topic worthy of in-depth investigation. Moreover, ML models, especially deep learning models, are data-driven, but the available data cannot provide sufficient information for ML models to achieve better results. Leveraging physical models to extract more meaningful information from limited data is a viable way to enhance model performance.

6. Conclusions

The poor generalizability is a common limitation across ML-based global ET models due to the sparse distribution of in situ observations. In this study, we developed a novel ensemble framework for combining the distinct advantages of process-based algorithms (extrapolation performance in data-sparse regions), ML-based ET models (data adaptability in data-dense regions), and hybrid models (overall performance) by introducing an additional ML classifier. Taking advantage of the ML classifier's capacity for automatic model selection, we are able to improve the generalizability of ML-based ET models by employing physically-founded process-based algorithm and hybrid model at appropriate pixel and avoid the typical underestimation of high values by ensemble methods.

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

The evaluation results across site and catchment scales indicate that our Classifier-Guided Ensemble model is overall more accurate than the individual base models, Attention-Based Ensemble model and other widely used global terrestrial ET products (ERA5, FLUXCOM, GLDAS, GLEAM_A, GLEAM_B, and PLSH) with lower RMSE and MAE and higher R^2 and KGE, especially in extreme samples where existing ML models perform poorly. At global scale, our model also exhibits higher stability, as well as greater consistency with these global ET products in both spatial patterns and latitude-averaged values.

In addition, the analysis of ML classifier's effectiveness demonstrates that the ML classifier can reasonably select the base models used at both global and site scales, highlighting the potential to further enhance the model's generalizability. Moreover, by further analyzing the SHAP values of different input covariates when the ML classifier select the 'dominant model', we gained a simple understanding of the mechanisms behind the selection of 'dominant model' that just as VPD, Ta, and Rs are critical for ET estimation, these variables also play a crucial role in model selection and identified some specific scenarios in which each model is most suitable. However, we chose to introduce global-scale training data to enhance the generalizability of the ML classifier, which has indeed led to improvements in ET estimation, but since these data were not obtained from ET observations, this may introduce uncertainties. Therefore, while our framework demonstrates significant potential to advance global ET estimation, further in-depth analysis and investigation are required, especially for the introduction of the ML classifier.

Appendix A: Supplementary tables

Table A1. Information for the 129 EC Flux Tower Sites, including the Site Name, Latitude (Lat), Longitude (Lon), International Geosphere-Biosphere Programme Land Cover Types (IGBP).

Name	Start year	End year	Lat	Long	IGBP
AR-SLu	2009	2011	-33.46	-66.46	MF
AT-Neu	2002	2012	47.12	11.32	GRA
AU-Ade	2007	2009	-13.08	131.12	SV
AU-ASM	2010	2014	-22.28	133.25	SV
AU-Cpr	2010	2014	-34.00	140.59	SV
AU-Cum	2012	2014	-33.62	150.72	EBF
AU-DaP	2007	2013	-14.06	131.32	GRA
AU-DaS	2008	2014	-14.16	131.39	SV
AU-Dry	2008	2014	-15.26	132.37	SV
AU-Emr	2011	2013	-23.86	148.47	GRA
AU-Fog	2006	2008	-12.55	131.31	WL
AU-Gin	2011	2014	-31.38	115.71	SV
AU-GWW	2013	2014	-30.19	120.65	SV

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

Name	Start year	End year	Lat	Long	IGBP
AU-How	2001	2014	-12.49	131.15	SV
AU-Lox	2008	2009	-34.47	140.66	DBF
AU-RDF	2011	2013	-14.56	132.48	SV
AU-Rig	2011	2014	-36.65	145.58	GRA
AU-Rob	2014	2014	-17.12	145.63	EBF
AU-Stp	2008	2014	-17.15	133.35	GRA
AU-TTE	2012	2014	-22.29	133.64	GRA
AU-Tum	2001	2014	-35.66	148.15	EBF
AU-Wac	2005	2008	-37.43	145.19	EBF
AU-Whr	2011	2014	-36.67	145.03	EBF
AU-Wom	2010	2014	-37.42	144.09	EBF
AU-Ync	2012	2014	-34.99	146.29	GRA
BE-Bra	1996	2014	51.31	4.52	MF
BE-Lon	2004	2014	50.55	4.75	CROP
BE-Vie	1996	2014	50.30	6.00	MF
BR-Sa3	2000	2004	-3.02	-54.97	EBF
CA-Gro	2003	2014	48.22	-82.16	MF
CA-Obs	1997	2010	53.99	-105.12	ENF
CA-Qfo	2003	2010	49.69	-74.34	ENF
CA-SF1	2003	2006	54.49	-105.82	ENF
CA-SF2	2001	2005	54.25	-105.88	ENF
CA-SF3	2001	2006	54.09	-106.01	OSH
CA-TP1	2002	2014	42.66	-80.56	ENF
CA-TP2	2002	2007	42.77	-80.46	ENF
CH-Cha	2005	2014	47.21	8.41	GRA
CH-Dav	1997	2014	46.82	9.86	ENF
CH-Fru	2005	2014	47.12	8.54	GRA
CN-Cng	2007	2010	44.59	123.51	GRA
CN-Du2	2006	2008	42.05	116.28	GRA
CN-Du3	2009	2010	42.06	116.28	GRA
CN-HaM	2002	2004	37.37	101.18	GRA
CZ-wet	2006	2014	49.02	14.77	WL
DE-Geb	2001	2014	51.10	10.91	CROP
DE-Gri	2004	2014	50.95	13.51	GRA
DE-Hai	2000	2012	51.08	10.45	DBF
DE-Kli	2004	2014	50.89	13.52	CROP

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

Name	Start year	End year	Lat	Long	IGBP
DE-Lkb	2009	2013	49.10	13.30	ENF
DE-Lnf	2002	2012	51.33	10.37	DBF
DE-Obe	2008	2014	50.79	13.72	ENF
DE-RuR	2011	2014	50.62	6.30	GRA
DE-SfN	2012	2014	47.81	11.33	WL
DE-Tha	1996	2014	50.96	13.57	ENF
DE-Zrk	2013	2014	53.88	12.89	WL
DK-Fou	2005	2005	56.48	9.59	CROP
DK-Sor	1996	2014	55.49	11.64	DBF
ES-LgS	2007	2009	37.10	-2.97	OSH
ES-LJu	2004	2013	36.93	-2.75	OSH
FI-Hyy	1996	2014	61.85	24.29	ENF
FI-Jok	2000	2003	60.90	23.51	CROP
FI-Let	2009	2012	60.64	23.96	ENF
FI-Lom	2007	2009	68.00	24.21	WL
FI-Sod	2001	2014	67.36	26.64	ENF
FR-LBr	1996	2008	44.72	-0.77	ENF
FR-Pue	2000	2014	43.74	3.60	EBF
GH-Ank	2011	2014	5.27	-2.69	EBF
GL-ZaF	2008	2011	74.48	-20.55	WL
IT-CA1	2011	2014	42.38	12.03	DBF
IT-CA2	2011	2014	42.38	12.03	CROP
IT-CA3	2011	2014	42.38	12.02	DBF
IT-Col	1996	2014	41.85	13.59	DBF
IT-Lav	2003	2014	45.96	11.28	ENF
IT-MBo	2003	2013	46.01	11.05	GRA
IT-PT1	2002	2004	45.20	9.06	DBF
IT-Ren	1998	2013	46.59	11.43	ENF
IT-Ro2	2002	2012	42.39	11.92	DBF
IT-Tor	2008	2014	45.84	7.58	GRA
MY-PSO	2003	2009	2.97	102.31	EBF
NL-Loo	1996	2014	52.17	5.74	ENF
RU-Cok	2003	2014	70.83	147.49	OSH
RU-Fyo	1998	2014	56.46	32.92	ENF
RU-Ha1	2002	2004	54.73	90.00	GRA
SD-Dem	2005	2009	13.28	30.48	SV

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

Name	Start year	End year	Lat	Long	IGBP
SN-Dhr	2010	2013	15.40	-15.43	SV
US-AR1	2009	2012	36.43	-99.42	GRA
US-AR2	2009	2012	36.64	-99.60	GRA
US-ARb	2005	2006	35.55	-98.04	GRA
US-ARc	2005	2006	35.55	-98.04	GRA
US-ARM	2003	2012	36.61	-97.49	CROP
US-Blo	1997	2007	38.90	-120.63	ENF
US-Cop	2001	2007	38.09	-109.39	GRA
US-CRT	2011	2013	41.63	-83.35	CROP
US-GBT	1999	2006	41.37	-106.24	ENF
US-GLE	2004	2014	41.37	-106.24	ENF
US-Goo	2002	2006	34.25	-89.87	GRA
US-Ivo	2004	2007	68.49	-155.75	WL
US-Lin	2009	2010	36.36	-119.09	CROP
US-Los	2000	2014	46.08	-89.98	WL
US-LWW	1997	1998	34.96	-97.98	GRA
US-Me1	2004	2005	44.58	-121.50	ENF
US-Me2	2002	2014	44.45	-121.56	ENF
US-Me4	1996	2000	44.50	-121.62	ENF
US-Me5	2000	2002	44.44	-121.57	ENF
US-MMS	1999	2014	39.32	-86.41	DBF
US-Ne1	2001	2013	41.17	-96.48	CROP
US-Ne2	2001	2013	41.16	-96.47	CROP
US-Ne3	2001	2013	41.18	-96.44	CROP
US-NR1	1998	2014	40.03	-105.55	ENF
US-Oho	2004	2013	41.55	-83.84	DBF
US-Prr	2010	2014	65.12	-147.49	ENF
US-SRC	2008	2014	31.91	-110.84	OSH
US-SRG	2008	2014	31.79	-110.83	GRA
US-SRM	2004	2014	31.82	-110.87	SV
US-Syv	2001	2014	46.24	-89.35	MF
US-Ton	2001	2014	38.43	-120.97	SV
US-Tw2	2012	2013	38.10	-121.64	CROP
US-Tw3	2013	2014	38.12	-121.65	CROP
US-Tw4	2013	2014	38.10	-121.64	WL
US-Twt	2009	2014	38.11	-121.65	CROP

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

Name	Start year	End year	Lat	Long	IGBP
US-Var	2000	2014	38.41	-120.95	GRA
US-WCr	1999	2014	45.81	-90.08	DBF
US-Whs	2007	2014	31.74	-110.05	OSH
US-Wi0	2002	2002	46.62	-91.08	ENF
US-Wi3	2002	2004	46.63	-91.10	DBF
US-Wi6	2002	2003	46.62	-91.30	OSH
US-Wkg	2004	2014	31.74	-109.94	GRA
ZM-Mon	2000	2009	-15.44	23.25	DBF

Table A2. Information for the 38 Selected Catchments.

Catchment Name	Continent	Catchment area ($\times 10^4$ km^2)
Amazon	South America	467.1
Congo	Africa	361.9
Ob	Asia	253.6
Parana Rio	South America	252.2
Yenisey	Asia	244.8
Lena	Asia	243.7
Yangtze	Asia	170.5
Mackenzie	North America	169.8
Volga	Europe	139.3
Missouri	North America	134.4
Orange	Africa	82.7
Yellow	Asia	73.0
Tocantins Rio	South America	69.7
South Atlantic-Gulf	North America	69.2
Niger	Africa	66.5
Arkansas-White-Red	North America	64.2
Columbia	North America	60.3
Songhua	Asia	52.8
Upper Mississippi	North America	49.2
Texas-Gulf	North America	46.4
Ohio	North America	42.2
California	North America	41.5
Pearl	Asia	41.5
Kolyma	Asia	37.1
Great Basin	North America	36.7

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

Catchment Name	Continent	Catchment area ($\times 10^4$ km^2)
Lower Colorado	North America	36.3
São Francisco	South America	34.5
Rio Grande	North America	34.3
Zambezi	Africa	33.5
Godavari	Asia	30.7
Parnaiba Rio	South America	29.8
Upper Colorado	North America	29.4
Lower Mississippi	North America	26.0
Mid-Atlantic	North America	25.2
Krishna	Asia	24.0
Cooper Creek	Australia	23.3
Okavango	Africa	22.9
Barwon	Australia	20.9

Table A3. Information for the 30 Selected EC Flux Tower Sites in Independent Validation.

Name	Start year	End year	Lat	Long	IGBP
AR-SLu	2009	2011	-33.46	-66.46	MF
AU-ASM	2010	2014	-22.28	133.25	SV
AU-Fog	2006	2008	-12.55	131.31	WL
AU-Ync	2012	2014	-34.99	146.29	GRA
BE-Bra	1996	2014	51.31	4.52	MF
BE-Vie	1996	2014	50.30	6.00	MF
BR-Sa3	2000	2004	-3.02	-54.97	EBF
CA-Qfo	2003	2010	49.69	-74.34	ENF
CA-SF3	2001	2006	54.09	-106.01	OSH
CH-Dav	1997	2014	46.82	9.86	ENF
CN-Du2	2006	2008	42.05	116.28	GRA
DE-Obe	2008	2014	50.79	13.72	ENF
DE-Tha	1996	2014	50.96	13.57	ENF
DK-Sor	1996	2014	55.49	11.64	DBF
ES-LgS	2007	2009	37.10	-2.97	OSH
FI-Hyy	1996	2014	61.85	24.29	ENF
FR-Pue	2000	2014	43.74	3.60	EBF
GH-Ank	2011	2014	5.27	-2.69	EBF
GL-ZaF	2008	2011	74.48	-20.55	WL
IT-CA1	2011	2014	42.38	12.03	DBF

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

Name	Start year	End year	Lat	Long	IGBP
RU-Cok	2003	2014	70.83	147.49	OSH
RU-Ha1	2002	2004	54.73	90.00	GRA
SD-Dem	2005	2009	13.28	30.48	SV
SN-Dhr	2010	2013	15.40	-15.43	SV
US-ARM	2003	2012	36.61	-97.49	CROP
US-Ivo	2004	2007	68.49	-155.75	WL
US-Ne1	2001	2013	41.17	-96.48	CROP
US-Oho	2004	2013	41.55	-83.84	DBF
US-Twt	2009	2014	38.11	-121.65	CROP
ZM-Mon	2000	2009	-15.44	23.25	DBF

Table A4. Performance of Different ET Models in Independent Validation.

Name	Classifier-Guided		Attention-Based		ML model		Hybrid model	
	R^2	KGE	R^2	KGE	R^2	KGE	R^2	KGE
AR-SLu	0.11	0.43	0.10	0.40	0.08	0.41	0.44	0.56
AU-ASM	0.85	0.68	0.78	0.66	0.78	0.61	0.76	0.65
AU-Fog	-0.40	0.17	-0.41	0.02	-0.44	0.05	-0.43	0.03
AU-Ync	-1.79	0.32	-2.47	0.20	-2.45	0.25	-2.21	0.24
BE-Bra	0.31	0.36	0.31	0.37	0.25	0.33	0.40	0.42
BE-Vie	0.86	0.78	0.84	0.80	0.84	0.78	0.82	0.78
BR-Sa3	-0.27	0.58	-0.02	0.53	-0.03	0.50	-0.84	0.57
CA-Qfo	0.89	0.85	0.87	0.85	0.86	0.81	0.88	0.85
CA-SF3	0.87	0.85	0.87	0.89	0.86	0.88	0.87	0.89
CH-Dav	0.15	0.57	0.01	0.53	0.08	0.53	-0.03	0.53
CN-Du2	0.89	0.87	0.87	0.86	0.87	0.86	0.86	0.86
DE-Obe	0.87	0.87	0.87	0.87	0.85	0.87	0.87	0.87
DE-Tha	0.76	0.69	0.75	0.67	0.75	0.68	0.75	0.69
DK-Sor	0.84	0.80	0.83	0.84	0.82	0.80	0.83	0.81
ES-LgS	-0.25	0.41	-0.59	0.30	-0.60	0.37	-0.86	0.22
FI-Hyy	0.87	0.82	0.83	0.78	0.83	0.79	0.85	0.81
FR-Pue	0.17	0.54	0.14	0.53	0.18	0.55	0.08	0.50
GH-Ank	-0.90	-0.04	-1.11	-0.07	-1.12	-0.07	-0.35	-0.08
GL-ZaF	-0.58	0.09	-0.66	0.07	-0.80	-0.04	-0.56	0.09
IT-CA1	0.67	0.59	0.70	0.63	0.69	0.63	0.68	0.59
RU-Cok	-0.56	0.63	-1.60	0.52	-0.80	0.61	-1.51	0.54
RU-Ha1	0.87	0.83	0.85	0.80	0.86	0.82	0.92	0.86

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

Name	Classifier-Guided		Attention-Based		ML model		Hybrid model	
	R^2	KGE	R^2	KGE	R^2	KGE	R^2	KGE
SD-Dem	0.11	0.15	0.09	0.14	0.11	0.15	0.04	0.12
SN-Dhr	0.74	0.70	0.64	0.64	0.71	0.67	0.58	0.61
US-ARM	0.66	0.83	0.44	0.73	0.46	0.74	0.45	0.74
US-Ivo	0.63	0.79	0.51	0.68	0.56	0.76	0.41	0.67
US-Ne1	0.94	0.85	0.94	0.87	0.93	0.84	0.94	0.84
US-Oho	0.77	0.62	0.76	0.62	0.75	0.60	0.74	0.59
US-Twt	0.73	0.68	0.70	0.75	0.69	0.68	0.70	0.66
ZM-Mon	0.68	0.83	0.66	0.77	0.64	0.78	0.65	0.80
Average	0.35	0.60	0.25	0.57	0.27	0.57	0.26	0.58

Data and Code availability

All dataset used are publicly available from data sources cited throughout the paper. The in-situ datasets were obtained from the FLUXNET2015 dataset via https://fluxnet.org/. The ERA5-Land reanalysis products were obtained from the ECMWF via https://cds.climate.copernicus.eu/. The water-balance-based evapotranspiration product was downloaded from the National Tibetan Plateau Data Center via https://doi.org/10.11888/Atmos.tpdc.300493. The FLUXCOM product was obtained via https://www.bgc-jena.mpg.de/geodb/projects/Home.php. The PLSH product was obtained via http://files.ntsg.umt.edu/data/. The GLEAM product version 3.8a and 3.8b was obtained via https://www.gleam.eu/. The GLDAS product was obtained via https://doi.org/10.5067/SXAVCZFAQLNO. The NDVI and LAI data was obtained from GIMMS product (Cao et al., 2023; Li et al., 2023). The MODIS Land Cover Climate Modeling Grid Product (MCD12C1) was obtained via https://doi.org/10.5067/MODIS/MCD12C1.006. The machine learning models were trained using AutoGluon version 0.8.2 (Erickson et al., 2020). The code supporting this study is available upon reasonable request from the corresponding author.

Author Contributions

Conceptualization: Le Ni, Weiguang Wang; Investigation: Le Ni, Weiguang Wang, Jianyu Fu, Mingzhu Cao; Supervision: Weiguang Wang; Visualization: Le Ni; Writing–original draft: Le Ni; Writing–review & editing: Weiguang Wang, Jianyu Fu, Mingzhu Cao.

Competing interests

The authors declare that they have no conflict of interest.

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

Acknowledgments

This work was jointly supported by the National Natural Science Foundation of China (U2240218, 52479010). J. F. is supported by the National Natural Science Foundation of China (42401020), National Key Laboratory of Water Disaster Prevention (2022nkms05). Cordial thanks are extended to the editor, the associate editor, and reviewers for their critical and constructive comments, which highly improve the quality of the manuscript.

References

Abbott, P. F. and Tabony, R. C.: The estimation of humidity parameters, Meteorological Magazine, 114, 49–56, 1985.

Allen, R. G., Pereira, L. S., Raes, D., Smith, M., and others: Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56, Fao, Rome, 300, D05109, 1998.

Ayan, E., Erbay, H., and Varçın, F.: Crop pest classification with a genetic algorithm-based weighted ensemble of deep convolutional neural networks, Computers and Electronics in Agriculture, 179, 105809, https://doi.org/10.1016/j.compag.2020.105809, 2020.

Bastiaanssen, W. G. M., Menenti, M., Feddes, R. A., and Holtslag, A. A. M.: A remote sensing surface energy balance algorithm for land (SEBAL). 1. Formulation, Journal of Hydrology, 212–213, 198–212, https://doi.org/10.1016/S0022-1694(98)00253-4, 1998.

Beaudoing, H., Rodell, M., and NASA/GSFC/HSL.: GLDAS Noah Land Surface Model L4 3 hourly 0.25 x 0.25 degree V2.1, Goddard Earth Sciences Data and Information Services Center (GES DISC) [data set], https://doi.org/10.5067/E7TYRXPJKWOQ, 2020.

Breiman, L.: Random Forests, Machine Learning, 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Brenowitz, N. D. and Bretherton, C. S.: Prognostic Validation of a Neural Network Unified Physics Parameterization, Geophysical Research Letters, 45, 6289–6298, https://doi.org/10.1029/2018GL078510, 2018.

Cai, Y., Xu, Q., Bai, F., Cao, X., Wei, Z., Lu, X., Wei, N., Yuan, H., Zhang, S., Liu, S., Zhang, Y., Li, X., and Dai, Y.: Reconciling Global Terrestrial Evapotranspiration Estimates From Multi-Product Intercomparison and Evaluation, Water Resources Research, 60, e2024WR037608, https://doi.org/10.1029/2024WR037608, 2024.

Cao, M., Wang, W., Xing, W., Wei, J., Chen, X., Li, J., and Shao, Q.: Multiple sources of uncertainties in satellite retrieval of terrestrial actual evapotranspiration, Journal of Hydrology, 601, 126642. https://doi.org/10.1016/j.jhydrol.2021.126642, 2021.

Cao, S., Li, M., Zhu, Z., Wang, Z., Zha, J., Zhao, W., Duanmu, Z., Chen, J., Zheng, Y., Chen, Y., Myneni, R. B., and Piao, S.: Spatiotemporally consistent global dataset of the GIMMS leaf area index (GIMMS LAI4g) from 1982 to 2020, Earth System Science Data, 15, 4877–4899, https://doi.org/10.5194/essd-15-4877-2023, 2023.

Chakraborty, D., Başağaoğlu, H., and Winterle, J.: Interpretable vs. noninterpretable machine learning models for data-driven hydro-climatological process modeling, Expert Systems with Applications, 170, 114498,

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

https://doi.org/10.1016/j.eswa.2020.114498, 2021.

Chen, H., Huang, J. J., Dash, S. S., Wei, Y., and Li, H.: A hybrid deep learning framework with physical process description for simulation of evapotranspiration, Journal of Hydrology, 606, 127422, https://doi.org/10.1016/j.jhydrol.2021.127422, 2022.

Chu, W., Zhang, C., Li, H., Zhang, L., Shen, D., and Li, R.: SHAP-powered insights into spatiotemporal effects: Unlocking explainable Bayesian-neural-network urban flood forecasting, International Journal of Applied Earth Observation and Geoinformation, 131, 103972, https://doi.org/10.1016/j.jag.2024.103972, 2024.

Dorogush, A. V., Ershov, V., and Gulin, A.: CatBoost: gradient boosting with categorical features support, CoRR, abs/1810.11363, 2018.

ElGhawi, R., Kraft, B., Reimers, C., Reichstein, M., Körner, M., Gentine, P., and Winkler, A. J.: Hybrid modeling of evapotranspiration: inferring stomatal and aerodynamic resistances using combined physics-based and machine learning, Environ. Res. Lett., 18, 034039, https://doi.org/10.1088/1748-9326/acbbe0, 2023.

Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., and Smola, A.: AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data, arXiv e-prints, arXiv:2003.06505, https://doi.org/10.48550/arXiv.2003.06505, 2020.

Eskandari, H., Saadatmand, H., Ramzan, M., and Mousapour, M.: Innovative framework for accurate and transparent forecasting of energy consumption: A fusion of feature selection and interpretable machine learning, Applied Energy, 366, 123314, https://doi.org/10.1016/j.apenergy.2024.123314, 2024.

Fan, J., Ma, X., Wu, L., Zhang, F., Yu, X., and Zeng, W.: Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data, Agricultural Water Management, 225, 105758, https://doi.org/10.1016/j.agwat.2019.105758, 2019.

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

Foken, T.: The Energy Balance Closure Problem: An Overview, Ecological Applications, 18, 1351–1367, https://doi.org/10.1890/06-0922.1, 2008.

Friedl, M. and Sulla-Menashe, D.: MCD12C1 MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 0.05Deg CMG V006, NASA Land Processes Distributed Active Archive Center [data set], https://doi.org/10.5067/MODIS/MCD12C1.006, 2015.

Fu, J., Wang, W., Shao, Q., Xing, W., Cao, M., Wei, J., Chen, Z., and Nie, W.: Improved global evapotranspiration estimates using proportionality hypothesis-based water balance constraints, Remote Sensing of Environment, 279, 113140, https://doi.org/10.1016/j.rse.2022.113140, 2022.

Galindo, F. J. and Palacio, J.: Estimating the Instabilities of N Correlated Clocks,

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

https://api.semanticscholar.org/CorpusID:92985051, 1999.

Galindo, F. J. and Palacio, J.: Post-processing ROA data clocks for optimal stability in the ensemble timescale, Metrologia, 40, S237, https://doi.org/10.1088/0026-1394/40/3/301, 2003.

Gan, R., Zhang, Y., Shi, H., Yang, Y., Eamus, D., Cheng, L., Chiew, F. H. S., and Yu, Q.: Use of satellite leaf area index estimating evapotranspiration and gross assimilation for Australian ecosystems, Ecohydrology, 11, e1974, https://doi.org/10.1002/eco.1974, 2018.

Ganaie, M. A., Hu, M., Malik, A. K., Tanveer, M., and Suganthan, P. N.: Ensemble deep learning: A review, Engineering Applications of Artificial Intelligence, 115, 105151, https://doi.org/10.1016/j.engappai.2022.105151, 2022.

Good, S. P., Noone, D., and Bowen, G.: Hydrologic connectivity constrains partitioning of global terrestrial water fluxes, Science, 349, 175–177, https://doi.org/10.1126/science.aaa5931, 2015.

Granata, F.: Evapotranspiration evaluation models based on machine learning algorithms—A comparative study, Agricultural Water Management, 217, 303–315, https://doi.org/10.1016/j.agwat.2019.03.015, 2019.

Greve, P. and Seneviratne, S. I.: Assessment of future changes in water availability and aridity, Geophysical Research Letters, 42, 5493–5499, https://doi.org/10.1002/2015GL064127, 2015.

Guo, X., Yao, Y., Tang, Q., Liang, S., Shao, C., Fisher, J. B., Chen, J., Jia, K., Zhang, X., Shang, K., Yang, J., Yu, R., Xie, Z., Liu, L., Ning, J., and Zhang, L.: Multimodel ensemble estimation of Landsat-like global terrestrial latent heat flux using a generalized deep CNN-LSTM integration algorithm, Agricultural and Forest Meteorology, 349, 109962, https://doi.org/10.1016/j.agrformet.2024.109962, 2024.

Huang, T. and Merwade, V.: Improving Bayesian Model Averaging for Ensemble Flood Modeling Using Multiple Markov Chains Monte Carlo Sampling, Water Resources Research, 59, e2023WR034947, https://doi.org/10.1029/2023WR034947, 2023.

Huntington, T. G.: Evidence for intensification of the global water cycle: Review and synthesis, Journal of Hydrology, 319, 83–95, https://doi.org/10.1016/j.jhydrol.2005.07.003, 2006.

Jiménez, C., Martens, B., Miralles, D. M., Fisher, J. B., Beck, H. E., and Fernández-Prieto, D.: Exploring the merging of the global land evaporation WACMOS-ET products based on local tower measurements, Hydrology and Earth System Sciences, 22, 4513–4533, https://doi.org/10.5194/hess-22-4513-2018, 2018.

Jung, M., Reichstein, M., Ciais, P., Seneviratne, S. I., Sheffield, J., Goulden, M. L., Bonan, G., Cescatti, A., Chen, J., de Jeu, R., Dolman, A. J., Eugster, W., Gerten, D., Gianelle, D., Gobron, N., Heinke, J., Kimball, J., Law, B. E., Montagnani, L., Mu, Q., Mueller, B., Oleson, K., Papale, D., Richardson, A. D., Roupsard, O., Running, S., Tomelleri, E., Viovy, N., Weber, U., Williams, C., Wood, E., Zaehle, S., and Zhang, K.: Recent decline in the global land evapotranspiration trend due to limited moisture supply, Nature, 467, 951–954, https://doi.org/10.1038/nature09396, 2010.

Jung, M., Reichstein, M., Margolis, H. A., Cescatti, A., Richardson, A. D., Arain, M. A., Arneth, A., Bernhofer, C., Bonal, D., Chen, J., Gianelle, D., Gobron, N., Kiely, G., Kutsch, W., Lasslop, G., Law, B. E., Lindroth, A., Merbold, L., Montagnani, L., Moors, E. J., Papale, D., Sottocornola, M., Vaccari, F., and Williams, C.: Global patterns of land-

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations, Journal of Geophysical Research: Biogeosciences, 116, https://doi.org/10.1029/2010JG001566, 2011.

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

Karbasi, M., Jamei, M., Ali, M., Malik, A., and Yaseen, Z. M.: Forecasting weekly reference evapotranspiration using Auto Encoder Decoder Bidirectional LSTM model hybridized with a Boruta-CatBoost input optimizer, Computers and Electronics in Agriculture, 198, 107121, https://doi.org/10.1016/j.compag.2022.107121, 2022.

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., and Kumar, V.: Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data, IEEE Transactions on Knowledge and Data Engineering, 29, 2318–2331, https://doi.org/10.1109/TKDE.2017.2720168, 2017.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y.: LightGBM: A Highly Efficient Gradient Boosting Decision Tree, in: Neural Information Processing Systems, 2017.

Koppa, A., Rains, D., Hulsman, P., Poyatos, R., and Miralles, D. G.: A deep learning-based hybrid model of global terrestrial evaporation, Nat Commun, 13, 1912, https://doi.org/10.1038/s41467-022-29543-7, 2022.

Kustas, W. P. and Norman, J. M.: A two-source approach for estimating turbulent fluxes using multiple angle thermal infrared observations, Water Resources Research, 33, 1495–1508, https://doi.org/10.1029/97WR00704, 1997.

Leuning, R., Zhang, Y. Q., Rajaud, A., Cleugh, H., and Tu, K.: A simple surface conductance model to estimate regional evaporation using MODIS leaf area index and the Penman-Monteith equation, Water Resources Research, 44, https://doi.org/10.1029/2007WR006562, 2008.

Li, C., Yang, H., Yang, W., Liu, Z., Jia, Y., Li, S., and Yang, D.: Error characterization of global land evapotranspiration products: Collocation-based approach, Journal of Hydrology, 612, 128102, https://doi.org/10.1016/j.jhydrol.2022.128102, 2022.

Li, M., Cao, S., Zhu, Z., Wang, Z., Myneni, R. B., and Piao, S.: Spatiotemporally consistent global dataset of the GIMMS Normalized Difference Vegetation Index (PKU GIMMS NDVI) from 1982 to 2022, Earth System Science Data, 15, 4181–4203, https://doi.org/10.5194/essd-15-4181-2023, 2023.

Liu, G., Tang, Z., Qin, H., Liu, S., Shen, Q., Qu, Y., and Zhou, J.: Short-term runoff prediction using deep learning multi-dimensional ensemble method, Journal of Hydrology, 609, 127762, https://doi.org/10.1016/j.jhydrol.2022.127762, 2022.

Liu, J., Chai, L., Dong, J., Zheng, D., Wigneron, J.-P., Liu, S., Zhou, J., Xu, T., Yang, S., Song, Y., Qu, Y., and Lu, Z.: Uncertainty analysis of eleven multisource soil moisture products in the third pole environment based on the three-corned hat method, Remote Sensing of Environment, 255, 112225, https://doi.org/10.1016/j.rse.2020.112225, 2021.

Lundberg, S. M. and Lee, S.-I.: A unified approach to interpreting model predictions, CoRR, abs/1705.07874, 2017.

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

Lyu, Y. and Yong, B.: A Novel Double Machine Learning Strategy for Producing High-Precision Multi-Source Merging Precipitation Estimates Over the Tibetan Plateau, Water Resources Research, 60, e2023WR035643, https://doi.org/10.1029/2023WR035643, 2024.

Ma, N., Zhang, Y., and Szilagyi, J.: Water-balance-based evapotranspiration for 56 large river basins: A benchmarking dataset for global terrestrial evapotranspiration modeling, Journal of Hydrology, 630, 130607, https://doi.org/10.1016/j.jhydrol.2024.130607, 2024.

MA N.: Dataset of the water-balance-based evapotranspiration of global typical large river basins (1983-2016), National Tibetan Plateau Data Center [data set], https://doi.org/10.11888/Atmos.tpdc.300493, 2024.

Mallick, K., Boegh, E., Trebs, I., Alfieri, J. G., Kustas, W. P., Prueger, J. H., Niyogi, D., Das, N., Drewry, D. T., Hoffmann, L., and Jarvis, A. J.: Reintroducing radiometric surface temperature into the Penman-Monteith formulation, Water Resources Research, 51, 6214–6243, https://doi.org/10.1002/2014WR016106, 2015.

Martens, B., Miralles, D. G., Lievens, H., van der Schalie, R., de Jeu, R. A. M., Fernández-Prieto, D., Beck, H. E., Dorigo, W. A., and Verhoest, N. E. C.: GLEAM v3: satellite-based land evaporation and root-zone soil moisture, Geoscientific Model Development, 10, 1903–1925, https://doi.org/10.5194/gmd-10-1903-2017, 2017.

Medlyn, B. E., Duursma, R. A., Eamus, D., Ellsworth, D. S., Prentice, I. C., Barton, C. V. M., Crous, K. Y., De Angelis, P., Freeman, M., and Wingate, L.: Reconciling the optimal and empirical approaches to modelling stomatal conductance, Global Change Biology, 17, 2134–2144, https://doi.org/10.1111/j.1365-2486.2010.02375.x, 2011.

Milly, P. C. D., Dunne, K. A., and Vecchia, A. V.: Global pattern of trends in streamflow and water availability in a changing climate, Nature, 438, 347–350, https://doi.org/10.1038/nature04312, 2005.

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

Miralles, D. G., Gentine, P., Seneviratne, S. I., and Teuling, A. J.: Land–atmospheric feedbacks during droughts and heatwaves: state of the science and current challenges, Annals of the New York Academy of Sciences, 1436, 19–35, https://doi.org/10.1111/nyas.13912, 2019.

Mizoguchi, Y., Miyata, A., Ohtani, Y., Hirata, R., and Yuta, S.: A review of tower flux observation sites in Asia, Journal of Forest Research, 14, 1–9, https://doi.org/10.1007/s10310-008-0101-9, 2009.

Mohammed, A. and Kora, R.: A comprehensive review on ensemble deep learning: Opportunities and challenges, Journal of King Saud University - Computer and Information Sciences, 35, 757–774, https://doi.org/10.1016/j.jksuci.2023.01.014, 2023.

Monteith, J. L.: Evaporation and environment., Symp Soc Exp Biol, 19, 205–234, 1965.

Mu, Q., Heinsch, F. A., Zhao, M., and Running, S. W.: Development of a global evapotranspiration algorithm based on MODIS and global meteorology data, Remote Sensing of Environment, 111, 519–536, https://doi.org/10.1016/j.rse.2007.04.015, 2007.

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

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

Muñoz Sabater: ERA5-Land monthly averaged data from 1950 to present, Copernicus Climate Change Service (C3S) Climate Data Store (CDS) [data set], https://doi.org/DOI: 10.24381/cds.68d2bb30, 2019.

Ochege, F. U., Yuan, X., Ezekwe, I. C., Ling, Q., Nzabarinda, V., Kayiranga, A., Xie, M., Shi, H., and Luo, G.: Reconstructing monthly 0.25° terrestrial evapotranspiration data in a remote arid region using Bayesian-driven ensemble learning method, Journal of Hydrology, 634, 131115, https://doi.org/10.1016/j.jhydrol.2024.131115, 2024.

Oki, T. and Kanae, S.: Global Hydrological Cycles and World Water Resources, Science, 313, 1068–1072, https://doi.org/10.1126/science.1128845, 2006.

de Oliveira e Lucas, P., Alves, M. A., de Lima e Silva, P. C., and Guimarães, F. G.: Reference evapotranspiration time series forecasting with ensemble of convolutional neural networks, Computers and Electronics in Agriculture, 177, 105700, https://doi.org/10.1016/j.compag.2020.105700, 2020.

Pascolini-Campbell, M. A., Reager, J. T., and Fisher, J. B.: GRACE-based Mass Conservation as a Validation Target for Basin-Scale Evapotranspiration in the Contiguous United States, Water Resources Research, 56, e2019WR026594, https://doi.org/10.1029/2019WR026594, 2020.

Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.-W., Poindexter, C., Chen, J., Elbashandy, A., Humphrey, M., Isaac, P., Polidori, D., Reichstein, M., Ribeca, A., van Ingen, C., Vuichard, N., Zhang, L., Amiro, B., Ammann, C., Arain, M. A., Ardö, J., Arkebauer, T., Arndt, S. K., Arriga, N., Aubinet, M., Aurela, M., Baldocchi, D., Barr, A., Beamesderfer, E., Marchesini, L. B., Bergeron, O., Beringer, J., Bernhofer, C., Berveiller, D., Billesbach, D., Black, T. A., Blanken, P. D., Bohrer, G., Boike, J., Bolstad, P. V., Bonal, D., Bonnefond, J.-M., Bowling, D. R., Bracho, R., Brodeur, J., Brümmer, C., Buchmann, N., Burban, B., Burns, S. P., Buysse, P., Cale, P., Cavagna, M., Cellier, P., Chen, S., Chini, I., Christensen, T. R., Cleverly, J., Collalti, A., Consalvo, C., Cook, B. D., Cook, D., Coursolle, C., Cremonese, E., Curtis, P. S., D'Andrea, E., da Rocha, H., Dai, X., Davis, K. J., Cinti, B. D., Grandcourt, A. de, Ligne, A. D., De Oliveira, R. C., Delpierre, N., Desai, A. R., Di Bella, C. M., Tommasi, P. di, Dolman, H., Domingo, F., Dong, G., Dore, S., Duce, P., Dufrêne, E., Dunn, A., Dušek, J., Eamus, D., Eichelmann, U., ElKhidir, H. A. M., Eugster, W., Ewenz, C. M., Ewers, B., Famulari, D., Fares, S., Feigenwinter, I., Feitz, A., Fensholt, R., Filippa, G., Fischer, M., Frank, J., Galvagno, M., et al.: The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data, Sci Data, 7, 225, https://doi.org/10.1038/s41597-020-0534-3, 2020.

Penman, H. L.: Natural evaporation from open water, bare soil and grass, Proc. R. Soc. Lond. A, 193, 120–145, https://doi.org/10.1098/rspa.1948.0037, 1948.

Pérez-Rodríguez, J., Fernández-Navarro, F., and Ashley, T.: Estimating ensemble weights for bagging regressors based on the mean–variance portfolio framework, Expert Systems with Applications, 229, 120462, https://doi.org/10.1016/j.eswa.2023.120462, 2023.

Polhamus, A., Fisher, J. B., and Tu, K. P.: What controls the error structure in evapotranspiration models?, Agricultural and

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

Forest Meteorology, 169, 12–24, https://doi.org/10.1016/j.agrformet.2012.10.002, 2013.

Priestley, C. H. B. and Taylor, R. J.: On the Assessment of Surface Heat Flux and Evaporation Using Large Scale Parameters, Monthly Weather Review, 100, 81–92, 1972.

Purdy, A. J., Fisher, J. B., Goulden, M. L., Colliander, A., Halverson, G., Tu, K., and Famiglietti, J. S.: SMAP soil moisture improves global evapotranspiration, Remote Sensing of Environment, 219, 1–14, https://doi.org/10.1016/j.rse.2018.09.023, 2018.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat: Deep learning and process understanding for data-driven Earth system science, Nature, 566, 195–204, https://doi.org/10.1038/s41586-019-0912-1, 2019.

Reitz, M., Sanford, W. E., and Saxe, S.: Ensemble Estimation of Historical Evapotranspiration for the Conterminous U.S., Water Resources Research, 59, e2022WR034012, https://doi.org/10.1029/2022WR034012, 2023.

Rienecker, M. M., Suarez, M. J., Gelaro, R., Todling, R., Bacmeister, J., Liu, E., Bosilovich, M. G., Schubert, S. D., Takacs, L., Kim, G.-K., Bloom, S., Chen, J., Collins, D., Conaty, A., Silva, A. da, Gu, W., Joiner, J., Koster, R. D., Lucchesi, R., Molod, A., Owens, T., Pawson, S., Pegion, P., Redder, C. R., Reichle, R., Robertson, F. R., Ruddick, A. G., Sienkiewicz, M., and Woollen, J.: MERRA: NASA's Modern-Era Retrospective Analysis for Research and Applications, https://doi.org/10.1175/JCLI-D-11-00015.1, 2011.

Rodell, M., Houser, P. R., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C.-J., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., Entin, J. K., Walker, J. P., Lohmann, D., and Toll, D.: The Global Land Data Assimilation System, Bulletin of the American Meteorological Society, 85, 381–394, https://doi.org/10.1175/BAMS-85-3-381, 2004.

Schwalm, C. R., Anderegg, W. R. L., Michalak, A. M., Fisher, J. B., Biondi, F., Koch, G., Litvak, M., Ogle, K., Shaw, J. D., Wolf, A., Huntzinger, D. N., Schaefer, K., Cook, R., Wei, Y., Fang, Y., Hayes, D., Huang, M., Jain, A., and Tian, H.: Global patterns of drought recovery, Nature, 548, 202–205, https://doi.org/10.1038/nature23021, 2017.

Shang, K., Yao, Y., Di, Z., Jia, K., Zhang, X., Fisher, J. B., Chen, J., Guo, X., Yang, J., Yu, R., Xie, Z., Liu, L., Ning, J., and Zhang, L.: Coupling physical constraints with machine learning for satellite-derived evapotranspiration of the Tibetan Plateau, Remote Sensing of Environment, 289, 113519, https://doi.org/10.1016/j.rse.2023.113519, 2023.

Shapley, L. S.: A Value for N-Person Games, RAND Corporation, Santa Monica, CA, https://doi.org/10.7249/P0295, 1952.

da Silva Júnior, J. C., Medeiros, V., Garrozi, C., Montenegro, A., and Gonçalves, G. E.: Random forest techniques for spatial interpolation of evapotranspiration data from Brazilian's Northeast, Computers and Electronics in Agriculture, 166, 105017, https://doi.org/10.1016/j.compag.2019.105017, 2019.

Su, Z.: The Surface Energy Balance System (SEBS) for estimation of turbulent heat fluxes, Hydrology and Earth System Sciences, 6, 85–100, https://doi.org/10.5194/hess-6-85-2002, 2002.

Tavella, P. and Premoli, A.: Estimating the Instabilities of N Clocks by Measuring Differences of their Readings, Metrologia, 30, 479, https://doi.org/10.1088/0026-1394/30/5/003, 1994.

Teuling, A. J., Hirschi, M., Ohmura, A., Wild, M., Reichstein, M., Ciais, P., Buchmann, N., Ammann, C., Montagnani, L.,

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

Richardson, A. D., Wohlfahrt, G., and Seneviratne, S. I.: A regional perspective on trends in continental evaporation, Geophysical Research Letters, 36, https://doi.org/10.1029/2008GL036584, 2009.

Torcaso, F., Ekstrom, C., Burt, E., and Matsakis, D.: Estimating Frequency Stability and Cross-Correlations, 14, 1998.

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.

Trenberth, K. E., Fasullo, J. T., and Kiehl, J.: Earth's Global Energy Budget, Bulletin of the American Meteorological Society, 90, 311–324, https://doi.org/10.1175/2008BAMS2634.1, 2009.

Tseng, M.-H.: GA-based weighted ensemble learning for multi-label aerial image classification using convolutional neural networks and vision transformers, Mach. Learn.: Sci. Technol., 4, 045045, https://doi.org/10.1088/2632-2153/ad10cf, 2023.

Twine, T. E., Kustas, W. P., Norman, J. M., Cook, D. R., Houser, P. R., Meyers, T. P., Prueger, J. H., Starks, P. J., and Wesely, M. L.: Correcting eddy-covariance flux underestimates over a grassland, Agricultural and Forest Meteorology, 103, 279–300, https://doi.org/10.1016/S0168-1923(00)00123-4, 2000.

Wang, D., and Alimohammadi, N.: Responses of annual runoff, evaporation, and storage change to climate variability at the watershed scale, Water Resources Research, 48(5), W05546. https://doi.org/10.1029/2011WR011444, 2012.

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

Wang, K., Dickinson, R. E., Wild, M., and Liang, S.: Evidence for decadal variation in global terrestrial evapotranspiration between 1982 and 2002: 1. Model development, Journal of Geophysical Research: Atmospheres, 115, https://doi.org/10.1029/2009JD013671, 2010a.

Wang, K., Dickinson, R. E., Wild, M., and Liang, S.: Evidence for decadal variation in global terrestrial evapotranspiration between 1982 and 2002: 2. Results, Journal of Geophysical Research: Atmospheres, 115, https://doi.org/10.1029/2010JD013847, 2010b.

Willard, J., Jia, X., Xu, S., Steinbach, M., and Kumar, V.: Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems, ACM Comput. Surv., 55, 66:1-66:37, https://doi.org/10.1145/3514228, 2022.

Williams, D. G., Cable, W., Hultine, K., Hoedjes, J. C. B., Yepez, E. A., Simonneaux, V., Er-Raki, S., Boulet, G., Bruin, H. A. R. de, Chehbouni, A., Hartogensis, O. K., and Timouk, F.: Evapotranspiration components determined by stable isotope, sap flow and eddy covariance techniques, Agricultural and Forest Meteorology, 125, 241–258, https://doi.org/10.1016/j.agrformet.2004.04.008, 2004.

Wilson, K. B., Hanson, P. J., Mulholland, P. J., Baldocchi, D. D., and Wullschleger, S. D.: A comparison of methods for determining forest evapotranspiration and its components: sap-flow, soil water budget, eddy covariance and catchment water balance, Agricultural and Forest Meteorology, 106, 153–168, https://doi.org/10.1016/S0168-1923(00)00199-4,

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

2001.

Xiao, J., Zhuang, Q., Baldocchi, D. D., Law, B. E., Richardson, A. D., Chen, J., Oren, R., Starr, G., Noormets, A., Ma, S., Verma, S. B., Wharton, S., Wofsy, S. C., Bolstad, P. V., Burns, S. P., Cook, D. R., Curtis, P. S., Drake, B. G., Falk, M., Fischer, M. L., Foster, D. R., Gu, L., Hadley, J. L., Hollinger, D. Y., Katul, G. G., Litvak, M., Martin, T. A., Matamala, R., McNulty, S., Meyers, T. P., Monson, R. K., Munger, J. W., Oechel, W. C., Paw U, K. T., Schmid, H. P., Scott, R. L., Sun, G., Suyker, A. E., and Torn, M. S.: Estimation of net ecosystem carbon exchange for the conterminous United States by combining MODIS and AmeriFlux data, Agricultural and Forest Meteorology, 148, 1827–1847, https://doi.org/10.1016/j.agrformet.2008.06.015, 2008.

Xie, X., Liang, S., Yao, Y., Jia, K., Meng, S., and Li, J.: Detection and attribution of changes in hydrological cycle over the Three-North region of China: Climate change versus afforestation effect, Agricultural and Forest Meteorology 203, 74–87. https://doi.org/10.1016/j.agrformet.2015.01.003, 2015.

Xie, Z., Yao, Y., Tang, Q., Liu, M., Fisher, J. B., Chen, J., Zhang, X., Jia, K., Li, Y., Shang, K., Jiang, B., Yang, J., Yu, R., Zhang, X., Guo, X., Liu, L., Ning, J., Fan, J., and Zhang, L.: Evaluation of seven satellite-based and two reanalysis global terrestrial evapotranspiration products, Journal of Hydrology, 630, 130649, https://doi.org/10.1016/j.jhydrol.2024.130649, 2024.

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

Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., Wang, J., Gao, J., and Zhang, L.: Deep learning in environmental remote sensing: Achievements and challenges, Remote Sensing of Environment, 241, 111716, https://doi.org/10.1016/j.rse.2020.111716, 2020.

Zhang, C., Brodylo, D., Rahman, M., Rahman, M. A., Douglas, T. A., and Comas, X.: Using an object-based machine learning ensemble approach to upscale evapotranspiration measured from eddy covariance towers in a subtropical wetland, Science of The Total Environment, 831, 154969, https://doi.org/10.1016/j.scitotenv.2022.154969, 2022.

Zhang, K., Kimball, J. S., Nemani, R. R., Running, S. W., Hong, Y., Gourley, J. J., and Yu, Z.: Vegetation Greening and Climate Change Promote Multidecadal Rises of Global Land Evapotranspiration, Sci Rep, 5, 15956, https://doi.org/10.1038/srep15956, 2015.

Zhang, K., Zhu, G., Ma, J., Yang, Y., Shang, S., and Gu, C.: Parameter Analysis and Estimates for the MODIS Evapotranspiration Algorithm and Multiscale Verification, Water Resources Research, 55, 2211–2231, https://doi.org/10.1029/2018WR023485, 2019.

Zhang, Z., Qin, H., Li, J., Liu, Y., Yao, L., Wang, Y., Wang, C., Pei, S., Li, P., and Zhou, J.: Operation rule extraction based on deep learning model with attention mechanism for wind-solar-hydro hybrid system under multiple uncertainties, Renewable Energy, 170, 92–106, https://doi.org/10.1016/j.renene.2021.01.115, 2021.

Zhangzhong, L., Gao, H., Zheng, W., Wu, J., Li, J., and Wang, D.: Development of an evapotranspiration estimation method

https://doi.org/10.5194/egusphere-2025-4782
Preprint. Discussion started: 2 December 2025

for lettuce via mobile phones using machine vision: Proof of concept, Agricultural Water Management, 275, 108003, https://doi.org/10.1016/j.agwat.2022.108003, 2023.

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

Zhu, W., Tian, S., Wei, J., Jia, S., and Song, Z.: Multi-scale evaluation of global evapotranspiration products derived from remote sensing images: Accuracy and uncertainty, Journal of Hydrology, 611, 127982, https://doi.org/10.1016/j.jhydrol.2022.127982, 2022.