A benchmark dataset for global evapotranspiration estimation based on FLUXNET2015 from 2000 to 2022

Wangyipu Li[1,2], Zhaoyuan Yao[1,2], Yifan Qu[1,2], Hanbo Yang[3], Yang Song[4], Lisheng Song[5], Lifeng Wu[6], Yaokui Cui[1,2]

[1]Institute of RS and GIS, School of Earth and Space Sciences, Peking University, Beijing 100871, China
[2]Beijing Key Laboratory of Spatial Information Integration and Its Applications, Beijing 100871, China
[3]State Key Laboratory of Hydro-science and Engineering, Department of Hydraulic Engineering, Tsinghua University, Beijing 100084, China
[4]Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China
[5]Key Laboratory of Earth Surface Processes and Regional Response in the Yangtze–Huaihe River Basin, School of Geography and Tourism, Anhui Normal University, Wuhu 241000, China
[6]Faculty of Modern Agricultural Engineering, Kunming University of Science and Technology, Kunming 650500, China

Correspondence to: Yaokui Cui (yaokuicui@pku.edu.cn)

Abstract. Evapotranspiration (ET) is a crucial component of terrestrial hydrological cycle. Latent heat flux (LE, equivalent to ET in W/m²) observed by the eddy covariance (EC) technique, commonly known as LE_{EC}, has been widely recognized as a highly accurate benchmark for global ET estimation. Currently, there is an increasing need for long time-series benchmark data to support climate change analysis, construction of new models, and validation of new products. However, existing LE_{EC} datasets, like FLUXNET2015, face significant challenges due to limited observation periods and extensive data gaps, which hinders their application in ET modelling and global change analysis. To address these issues, we developed a gap-filling and prolongation framework for LE_{EC} data and established a benchmark dataset for global ET estimation from 2000 to 2022 across 64 sites at various time scales. The framework mainly includes three parts: site selection and data pre-processing, gap-filled half-hourly/hourly LE data generation, and prolonged daily LE data generation. We selected 64 sites from FLUXNET2015 based on rigorous filtering criteria. A novel bias-corrected random forest (RF) algorithm was used as the gap-filling and prolongation algorithm in the framework to produce seamless half-hourly and daily LE data. After analysis, the framework using the novel bias-corrected RF algorithm achieves excellent performance in both hourly gap-filling and daily prolongation, with mean RMSE values of 33.86 W/m^2 and 16.58 W/m^2, respectively. The algorithm significantly improved the gap-filling performance for long gaps and extreme values compared with the original RF and marginal distribution sampling (MDS) algorithm. The results demonstrate robust prolongation performance of our framework in both prolongation directions and temporal stability. There is high consistency in data distribution between our gap-filled dataset and the FLUXNET2015 dataset. In conclusion, we have published the first benchmark dataset for global ET estimation based on FLUXNET2015 from 2000 to 2022. This dataset can effectively provide data support for ET modelling, water-carbon cycle monitoring, and climate change analysis. It is made freely available via the following repository: https://doi.org/10.5281/zenodo.13853409 (Li et al., 2024b).

1 Introduction

Terrestrial evapotranspiration (ET), which represents the movement and phase change of water from land to the atmosphere, is the second most critical component of the hydrological cycle (Zhang et al., 2016; Cui et al., 2021a; Yang et al., 2023; Song et al., 2024; Tang et al., 2024). It accounts for more than 60% of the land surface water derived from precipitation that returns to the atmosphere (Oki and Kanae, 2006). Therefore, it is essential to accurately estimate the magnitude and variability of global ET.

Ground-based instruments for observing ET are widely distributed globally. The eddy covariance technique is the most commonly used method, providing high-frequency (10–20 Hz) measurements of vertical wind speed and water vapor density (Aubinet et al., 2012; Pastorello et al., 2020). By calculating their covariance, the latent heat flux (LE, equivalent to ET in W/m^2; hereafter LE is used when describing ground observations) is derived. The EC technique offers several advantages, including non-destructive measurement of the underlying surface environment and flexible installation (Baldocchi, 2020; Pastorello et al., 2020). However, challenges remain in practical applications when LE obtained from the EC technique (LE_{EC}) primarily serves the two research communities:

(1) The global change analysis research community. With the abundance of remotely-sensed and reanalysis data and the development of ET models, more and more ET products based on remote sensing or earth system model simulation are produced and shared (Mu et al., 2011; Martens et al., 2017; Zhang et al., 2019; Cui and Jia, 2021; Zheng et al., 2022). However, their results differ significantly in average annual totals, temporal trends, and spatial distribution, which prevents us from properly understanding current changes in ET and the water-carbon cycle (Chen et al., 2014; Hu et al., 2021; Cui et al., 2023; Yang et al., 2023; Tang et al., 2024). Since LE_{EC} data are considered as the ground truth, researchers are eager to find evidence from ground observations to support their hypotheses. As the most widely used LE_{EC} dataset, the FLUXNET2015 dataset only provides observations up to 2014 (Pastorello et al., 2020). It cannot support global climate change analysis, nor can it help resolve discrepancies between different products;

(2) ET modelling community. First of all, many ET models (such as A-OPTRAM, PML-V2, and ETMonitor) require LE_{EC} data for parameter calibration to improve their performance (Zhang et al., 2019; Zheng et al., 2022; Yao et al., 2024). Second, all ET products must undergo validation by comparing themselves to LE_{EC} data (Mu et al., 2011; Zhang et al., 2016; Zhang et al., 2019; Cui et al., 2021b; Zheng et al., 2022). Especially for the latest models developed using new satellite data (such as SMAP, launched in 2015), there is a need to develop and validate them based on the latest ground-based benchmark data (Das et al., 2018; Zhang et al., 2024). However, due to limitations such as data sharing policies, the research community still relies on FLUXNET2015 as the primary source for calibration and validation. With the acceleration of the global water and energy cycle, parameters calibrated using outdated data may no longer be applicable, and it is difficult to assess model performance over the past decade. The research community aspires to leverage the most recent, long-term LE_{EC} data; however, there is a lack of up-to-date datasets that are readily accessible for their use.

Therefore, the two main issues with LE_{EC} data, such as those represented by FLUXNET2015, are:

(1) Extensive data gaps. There is a substantial amount of missing data in LE_{EC}. The missing rate of hourly data is around 40% and can be up to 70% for some sites. Long gaps, such as the 30-day gap scenario, account for an average of 44% of all missing data in FLUXNET2015. Although the marginal distribution sampling (MDS) algorithm is used as the official gap-filling algorithm, its performance in filling these long gaps is suboptimal (Foltýnová et al., 2020; Zhu et al., 2022);

(2) Limited observation duration. Only 33% of the sites have observation periods exceeding 10 years, and few sites have more than 20 years of observations. After quality control, less than half of the sites have observation periods longer than 8 years. The MDS algorithm can only be used for gap-filling but not for data prolongation. The potential of LE_{EC} data is not fully exploited. Therefore, there is an urgent need for a long-term ET benchmark dataset based on ground observations with temporally continuous and high-quality data.

To address this, we developed a gap-filling and prolongation framework for LE_{EC} data and established a benchmark dataset for global ET estimation from 2000 to 2022 across 64 sites. We selected 64 sites out of 206 public sites from FLUXNET2015 based on rigorous filtering criteria. After pre-processing the reanalysis and remote sensing data, the time series data of reference variables for each station are obtained. Then, a novel bias-corrected random forest (RF) algorithm was used as the core method of the framework to produce seamless half-hourly and daily LE data. We designed comprehensive experiments to evaluate our results, including assessing performance under different gap-length scenarios for gap-filling, evaluating consistency between forward and backward prolongation, and analysing the temporal stability of the prolongation data. This dataset aims to provide valuable data support for global ET modelling, water-carbon cycle monitoring, and climate change research.

2 Data

2.1 FLUXNET2015

The FLUXNET2015 dataset contains land-atmosphere exchange data of energy and carbon from 212 globally distributed sites (206 sites under CC-BY 4.0 license) (https://fluxnet.org/data/fluxnet2015-dataset/). We primarily used the LE data observed by EC technique and some auxiliary meteorological observations. From the original measurements to the hourly/half-hourly products, both datasets underwent a strict and uniform processing procedure across all sites, with additional scrutiny for these critical variables (Pastorello et al., 2020). After quality assurance and quality control (QA/QC), data that did not meet the standards or were missing due to power failures or sensor malfunctions were filtered out and QC flags were given. Only data marked as "0" was regarded as reliable ground observations, while other data were gap-filled by the MDS algorithm, with confidence levels decreasing as the flag number increased. For our analysis, we exclusively used LE and meteorological data marked as "0".

2.2 ERA5-Land

We used the latest Reanalysis v5 dataset (ERA5-Land) provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) (Muñoz-Sabater et al., 2021) as the source of reference data (https://www.ecmwf.int/en/era5-land). This dataset provides globally seamless meteorological data at a spatiotemprol resolution of 0.1°× 0.1° and 1 hour since 1950. The dataset provided meteorological variables including air temperature (TA), u-component of wind (u), v-component of wind (v), dewpoint temperature (TD), incoming shortwave radiation (SW_IN), incoming longwave radiatio (LW_IN), and air pressure (PA). Wind speed (WS) was calculated by the two component (u and v), and relative humidity (RH) was caculated by the following equations:

$$RH = \frac{e}{e_s} \times 100\% , \quad (1)$$

$$e_s = 6.1078 \times \exp\left(\frac{a \times T_a}{Ta + 273.15 - b}\right) \begin{cases} a = 17.27,\ b = 35.86, T_a > 0 \\ a = 21.87,\ b = 7.66, T_a \le 0 \end{cases}, \quad (2)$$

$$e = 6.1078 \times \exp\left(\frac{a \times T_d}{T_d + 273.15 - b}\right) \begin{cases} a = 17.27,\ b = 35.86, T_a > 0 \\ a = 21.87,\ b = 7.66, T_a \le 0 \end{cases}, \quad (3)$$

where e_s is the saturated vapour pressure (kPa), e is the actual vapour pressure (kPa), T_a is the air temperature, and T_d is the dewpoint temperature (°C).

2.3 MODIS

We obtained remotely-sensed normalized difference vegetation index (NDVI) data from the Moderate Resolution Imaging Spectroradiometer (MODIS) MYD13Q1.061 dataset. Its spatial and temporal resolutions are 250 m and 16 days, respectively. This dataset has been proven to be one of the most reliable NDVI datasets and is widely used in ET modelling.

3 Methodology

The gap-filling and prolongation framework for LE_{EC} data mainly includes 3 parts: site selection and data pre-processing, generation of gap-filled half-hourly or hourly LE data, and generation of prolonged daily LE data (Fig 1). The details are as follows:

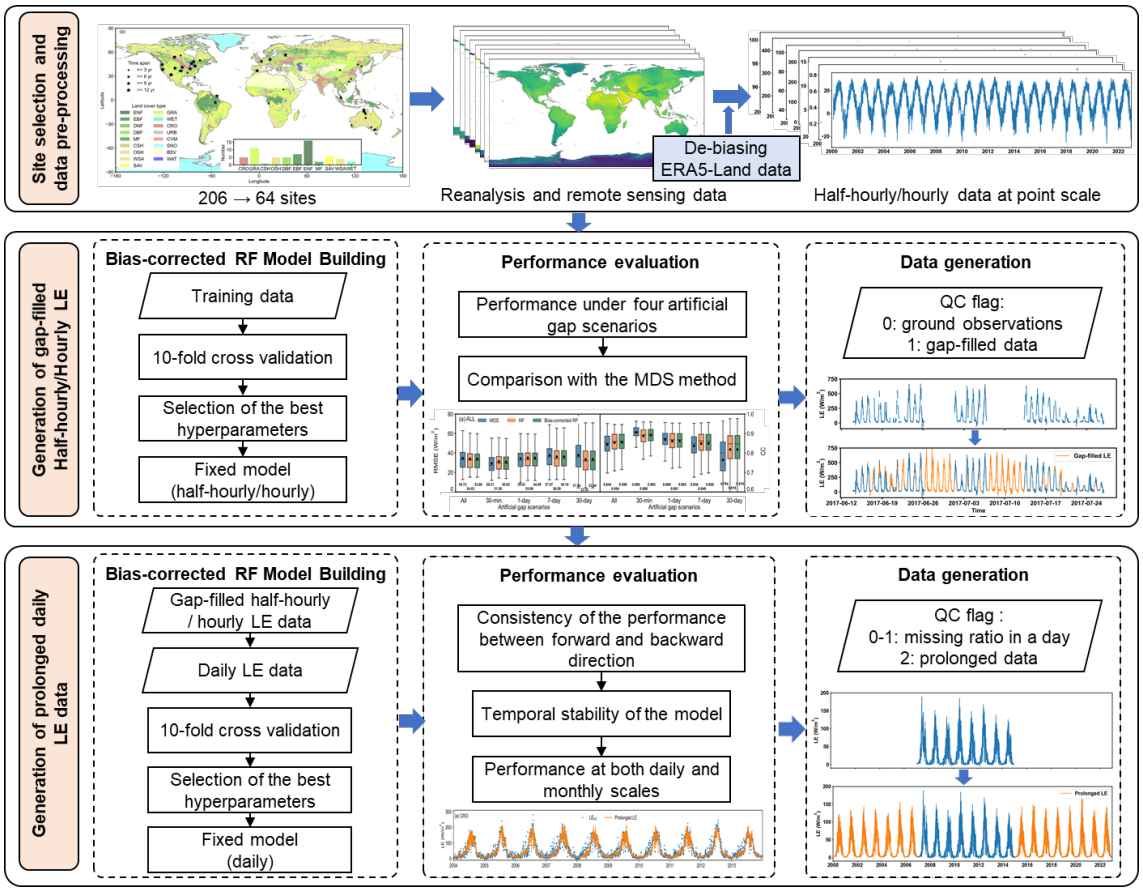


Figure 1 Schematic of the gap-filling and prolongation framework for LE_{EC} data.

3.1 Site selection and data pre-processing

3.1.1 FLUXNET2015 site selection

We selected 64 sites from 206 open-access FLUXNET2015 sites based on the following filtering criteria: 1) Time span. Sites must have ≥3 years of observations because sufficient temporal coverage is essential for reliable data prolongation; 2) Data missing rate. Sites must have ≤50% missing data, so that there are adequate data availability for half-hourly or hourly gap-filling; 3) Energy balance closure. We calculated the daily energy balance ratio (EBR) when there were ≥36 (18 for hourly data) valid observations in a day. The EBR values closest to 1 indicate the best agreement with the first law of thermodynamics, reflecting higher quality surface energy data. Sites were retained for analysis only if more than 20% of their observation days exhibited EBR values between 0.8 and 1.2. The EBR was calculated as follows:

$$EBR = \frac{\sum_{i=1}^{n}(LE+H)}{\sum_{i=1}^{n}(R_n-G)}, \quad (4)$$

where: R_n, G, and H are the net radiation, soil heat flux, and sensible heat flux, respectively.

Notably, no sites in Africa fully met the specified criteria. Consequently, we selected two additional sites that substantially met the essential requirements. In total, 64 sites were selected (Fig 2). These sites cover most regions, with 49 sites in the Northern Hemisphere, and 15 sites in the Southern Hemisphere. Sites in Europe and the Americas have longer observation periods, while those in Asia and Oceania are shorter. The average duration of observations across all sites is approximately 8 years. Between 2000 and 2014, observations were available from approximately 10 to 40 sites per year. Moreover, these sites represent the majority of vegetated land cover types. For detailed site information, see Table A1.

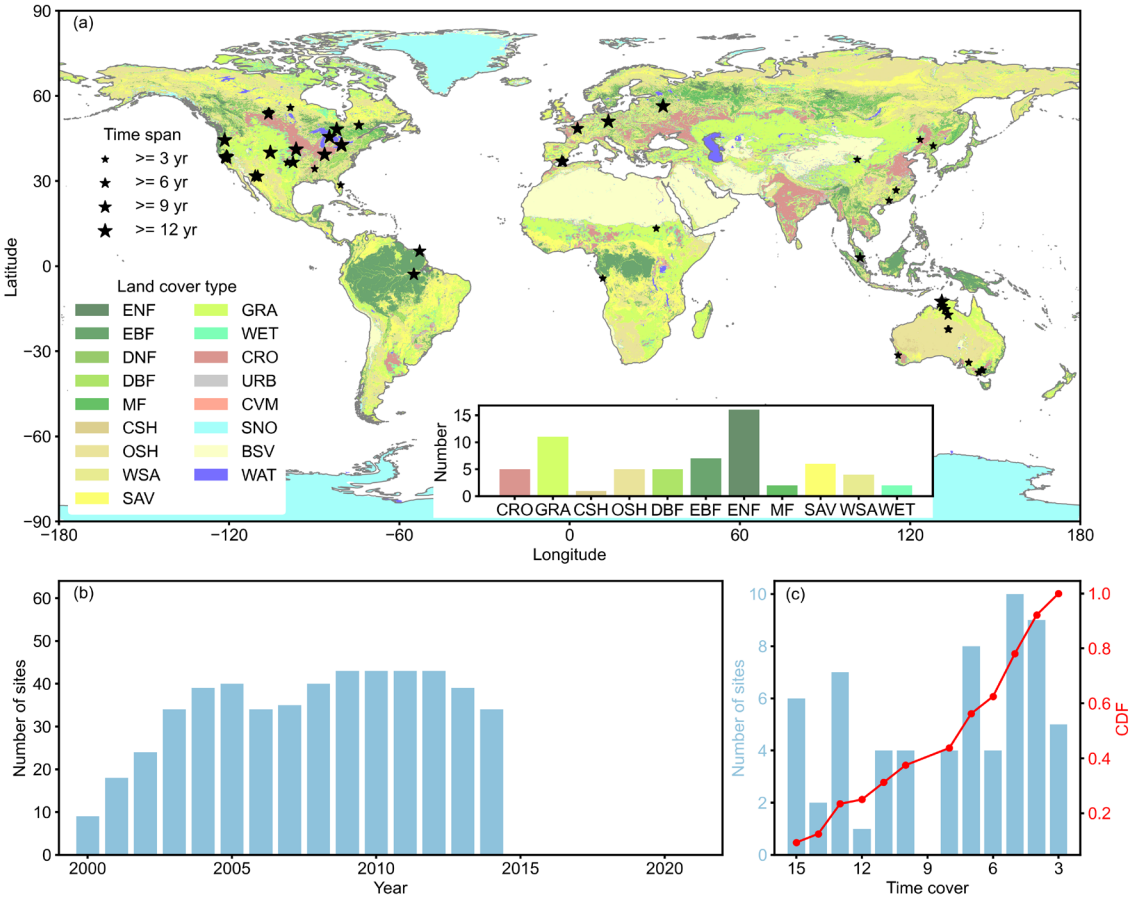


Figure 2 Global distribution (a) and information (b and c) of 64 selected FLUXNET2015 sites. The size of the star mark indicates the length of the data record. Panel (b) shows the number of sites in a year from 2000 to 2022. Panel (c) is the statistic of the length of observation periods for all sites.

3.1.2 Data pre-processing

We followed the same data pre-processing procedure as Li et al (2024). For the LE_{EC} data, non-observed values were filtered out based on quality control flags, with the remaining data used for training and testing datasets. The LE_{EC} data are reported at local time.

Reference variables, including TA, WS, RH, PA, SW_IN, LW_IN, and NDVI, were selected based on the Penman-Monteith (PM) equation (Monteith, 1965). These variables directly or indirectly influence the parameters in the PM equation and represent the most suitable variables to characterize meteorological and vegetated conditions affecting the ET process (Zhang et al., 2008; Mu et al., 2011; Li et al., 2024a). The PM equation is expressed as:

$$LE = \frac{\Delta R_n^* + \frac{\rho c_p \times VPD}{r_a}}{\Delta + \gamma\left(1 + \frac{r_s}{r_a}\right)}, \quad (5)$$

$$VPD = e_s - e, \quad (6)$$

where Δ *is* the slope of the vapor pressure curve, R_n^* is the net available radiation at the evaporating surface, ρ is the density of air, c_p is the specific heat of air at constant pressure, *VPD* is the air vapour pressure deficit, γ is a psychrometric constant, r_s is the surface resistance, and r_a is the aerodynamic resistance.

Reference variables from ERA5-Land and MODIS were extracted as time-series data at point scale using Google Earth Engine (https://code.earthengine.google.com/). Depending on the temporal resolution of LE_{EC} data records, the hourly time-series data from ERA5-Land were resampled to a half-hourly scale using linear interpolation or maintained at hourly scale. All timestamps were converted from UTC to local time to match the site-specific time zone. The NDVI data with a 16-day temporal resolution were resampled to a daily frequency using Savitzky-Golay filtering. The same value was then assigned uniformly for each day.

3.1.3 De-biasing the ERA5-Land data

To minimize mismatches between in-situ and raster data, the time-series data from ERA5-Land were further processed. We followed a procedure similar to the official products (Vuichard and Papale, 2015) and corrected biases between ground observations and ERA5-Land using a linear correction method:

$$Ground_i = k_i \times EL5_i + b_i, \quad (7)$$

where: i means different variables, EL5 is the ERA5-Land data, and Ground is the ground observations from FLUXNET2015. These variables were filtered by quality control flags and only valid observations were used. The ground observed vapour variable was VPD instead of RH in some sites. We transferred it to RH using the following equation:

$$RH = \left(1 - \frac{VPD}{e_s}\right) \times 100\%, \quad (8)$$

3.2 Generation of gap-filled half-hourly or hourly LE data

3.2.1 Bias-corrected random forest algorithm

Random Forest (RF), used for both classification and regression tasks, is composed of multiple decision trees, and it combines their predictions to generate the final output (Breiman, 2001). Numerous studies have demonstrated the

effectiveness of machine learning algorithms for gap-filling ground-based ET data (Moffat et al., 2007; Irvin et al., 2021; Mahabbati et al., 2021; Zhu et al., 2022; Li et al., 2024a). The RF algorithm is considered as one of the most robust and efficient machine learning algorithms for replacing the traditional MDS algorithm and has significant potential for prolonging time series. However, there has been limited research on prolonging LE_{EC} time series using RF, and no corresponding datasets have been released. Although the performance of RF in flux data gap-filling has been proven to be efficient, it still faces challenges, such as overestimating lower values and underestimating higher values. Therefore, it is necessary to correct the bias. Here, we chose a novel bias-corrected RF algorithm (Zhang and Lu, 2012). It added a bias correction RF model to improve the performance compared with original RF (Fig 3). This algorithm has been used for studies such as drought monitoring (Feng et al., 2019; Wang et al., 2023). In this study, the bias-corrected RF model was adapted for processing flux data, and the detailed procedure of this bias-correction method is summarized in Fig. 3.

In the model training step, we trained one model (including RF Model 1 and RF Model 2) for each site, resulting in a total of 64 models for data gap-filling task. For each site, the data were randomly divided into two parts: the training dataset (80% of the total dataset) and the test set (the remaining 20%). To optimize model performance and avoid overfitting, we employed a 10-fold cross-validation method to determine the optimal combination of hyperparameters. For each site, the training and test dataset were generated 20 times, so we did the 10-fold cross validation for 20 times and gained 20 hyperparameter combinations. We found that for each site, the 20 hyperparameter combinations are almost the same. Therefore, we choose the hyperparameter combination based on two criteria: (1) achieving optimal model performance, and (2) exhibiting the highest frequency of occurrence across 20 experimental trials. Consequently, each site has a site-specific and unique hyperparameter combination. Finally, we used all valid LE observations to build the model. See 3.2.2 for details on how to split the training and test sets.

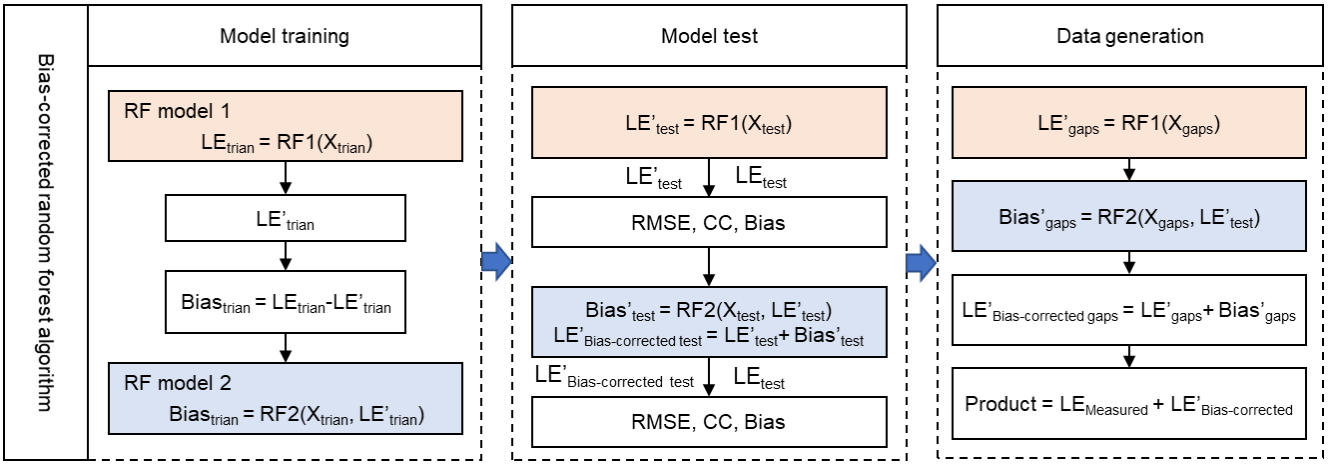


Figure 3 Schematic diagram of the bias-corrected RF algorithm. Train in the subscript indicates the training data. Test in the subscript indicates the test data. Gaps in the subscript indicates the data gap to be filled. LE and Bias with single quotes indicate predicted values, whereas those without single quotes indicate ground observations. X indicates the reference

variables, including TA, WS, RH, PA, SW_IN, LW_IN, and NDVI. Prolonging daily data also has the same processing steps.

3.2.2 Artificial gap scenarios

The length of gaps in LE_{EC} data varied significantly, ranging from one single missing record to gaps exceeding 30 days. To fully evaluate the performance of our model, we generated four gap-length scenarios , covering short to long durations: 30-minute, 1-day, 7-day and 30-day scenarios (Zhu et al., 2022; Li et al., 2024a). The artificial gaps for each scenario accounted for approximately 5% of the total dataset, and all gaps collectively constituted the test set (20%). After removing these artificial gaps, the remaining data (80%) were used to train the model. Specifically, we used a sliding window approach to generate gap scenarios. If the valid observed data coverage within a window exceeded 50%, the window was marked. The window automatically moved forward until this criterion was met, and no overlaps among marked sliding windows were ensued. The sliding window size was initially set to 30 days. After completing one full round of marking, we randomly selected gaps accounting for 5% of the total dataset, and these data were removed. The sliding window size was then reduced to 7 days and 1 day, and the process was repeated. Finally, we randomly removed 5% of half-hourly data to create the 30-min scenarios, ensuring 5 consecutive valid data points before and after each gap. To ensure result robustness, we generated 20 different training-test sets and repeated the above steps 20 times for each site.

For intercomparison, we also used the MDS algorithm and original RF algorithm as the gap-filling algorithm of the framework. The core of the MDS algorithm is to use a sliding window approach to find similar meteorological conditions (Reichstein et al., 2005). It primarily uses SW_IN, VPD, and TA as reference variables. The larger the sliding window, the lower the confidence in the gap-filling results. To closely simulate the official data producing process, this study set the minimum thresholds for the three variables at 50 W/m^2, 5 hPa, and 2.5°C, respectively. The MDS algorithm was implemented using the REddyProc (R package, v.1.3.3).

3.3 Generation of prolonged daily LE data

Current mainstream ET products are predominantly available at daily and monthly scales(Zhang et al., 2019; Zheng et al., 2022; Miralles et al., 2025). Therefore, prolonging daily-scale ET data aligns best with current practical application scenarios.

3.3.1 Data generation

Following half-hourly/hourly gap-filling, we obtained continuous time-series data. These data were then aggregated from half-hourly/hourly to daily resolution, with the daily missing rate included as a QC flag. For data prolongation, we employed the bias-corrected Random Forest (RF) algorithm, maintaining the same model architecture and training procedure described in Section 3.2.1. During model training, we trained one model for each site, selecting all data except for those with a missing rate of 1 (completely missing) for model training. The 10-fold cross-validation method was used to determine the optimal hyperparameters. Ultimately, the seamless daily LE data from 2000 to 2022 were produced. The final product has been deposited at https://doi.org/10.5281/zenodo.13853409 (Li et al., 2024b) and can be downloaded publicly.

3.3.2 Experimental design for evaluating the prolonged data

Since the number of days with a missing ratio of 0 at the daily scale is extremely limited, we considered that daily data with a missing ratio below 10% could serve as the test data.

The prolongation at the daily scale was conducted into two time directions: forward and backward. For example, one site has the time cover from 2007-2014. Therefore, prolongation of 2000-2006 is the backward direction, and prolongation of 2015-2022 is the forward direction. We expect that the prolongation performance will be consistent in both directions. To validate the consistency of our method, we adopted the following approach: backward prolongation, the first 1/3 of the data served as the test set, while the remaining 2/3 was used for training; for forward prolongation, the first 2/3 of the data was used for training, and the last 1/3 served as the test set. We then compared the performance of both directions.

To assess the temporal stability of the model's performance as the prolongation period increased, we conducted experiments using two representative observation lengths: for sites with ≥8 years of observations, we used the first 8 years as the training set and each subsequent year as the test set. For sites with ≥3 years of observations, we used the first 3 years as the training set and each subsequent year as the test set.

3.4 Performance metrics

We selected four commonly used performance metrics, including the root mean square error (RMSE, W/m^2), bias (Bias, W/m^2), correlation coefficient (CC), and coefficient of variation (CV). The equations are as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2}\,, \tag{9}$$

$$\mathrm{Bias} = \frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)\,, \tag{10}$$

$$\mathrm{CC} = \frac{\sum_{i=1}^{n}(p_i-\bar{p})(o_i-\bar{o})}{\sqrt{\sum_{i=1}^{n}(p_i-\bar{p})^2\sum_{i=1}^{n}(o_i-\bar{o})^2}}\,, \tag{11}$$

$$\mathrm{CV} = \frac{\sigma}{\mu}\times 100\%\,, \tag{12}$$

where p_i and o_i are the values from prediction and observation, respectively. $\bar{p}$ denotes the mean predicted value. σ is the standard deviation of the target data. μ is the averaged value of the target data.

4 Results

4.1 Evaluation of half-hourly or hourly gap-filled LE data

4.1.1 Gap-filling performance under different gap-length scenarios

We conducted a comprehensive evaluation of the gap-filling performance for three algorithms under artificially constructed gap scenarios, including the official algorithm (MDS), widely-used RF algorithm, and novel bias-corrected RF algorithm.

For each station and each combination of training and test set, we calculated the statistical metrics RMSE, CC, and Bias, and then visualized the results using box plots (Fig 4 and Fig 5).

In general, the results indicate that the gap-filled data obtained using the bias-corrected RF are superior to the official (MDS) algorithm, particularly outperforming it significantly for long gaps. The bias-corrected RF exhibits the best performance (33.86 W/m^2 and 0.86 in terms of mean RMSE and CC), with mean RMSE improvements of 0.87 W/m^2 and 0.16 W/m^2 compared to MDS and RF, respectively. As for the bias metric, Figure 5 shows that as the length of the gap length increases, the uncertainty increases and the bias-corrected RF provides more robust results.

For short gaps, we find that the performance of the bias-corrected RF is closer to those of the MDS compared to the original RF. Specifically, the MDS performs exceptionally well, with mean values of RMSE and CC at 29.31 W/m^2 and 0.91, respectively. The original RF performs the worst, while the bias-corrected RF reduce a bias (0.45 W/m^2 in terms of RMSE), making its performance closer to the MDS compared with the original RF. However, as the gap length increases, the performance of the MDS declines sharply, which is consistent with previous studies (Foltýnová et al., 2020; Zhu et al., 2022; Li et al., 2024a). Under the 30-day gap length scenario, the mean RMSE of the MDS (37.64 W/m^2) is 10.93% and 11.15% lower than those of RF (33.52 W/m^2) and bias-corrected RF (33.44 W/m^2), respectively. Due to the use of the sliding window method, MDS encounters significant issues during the early months of observation. Specifically, when data is completely missing for the early months, the results from MDS at the monthly scale are nearly the same value. A detailed analysis of this issue can be found in Section 5.1.

We further analyzed the gap-filling results across different land cover types. Based on station count, land cover characteristics, and relevant practices from previous studies, we categorized the land surface types into four groups for analysis: CRO, GRA, DBF/EBF/ENF/MF, and CSH/OSH/SAV/WSA/WET. Overall, for all land surface types, the bias-corrected RF performs better than the original RF and provides closer performance to MDS. Specifically, the bias-corrected RF shows the most significant improvement in CRO, with the mean RMSE being 4.26% lower compared to MDS. This indicates that incorporating NDVI as a reference variable can better capture the seasonal dynamics of crops. We also observed that in GRA and CSH/OSH/SAV/WSA/WET, the bias-corrected RF provides results closer to the gap-filling performance of MDS and the MDS performs much better than the original RF. Across different gap length scenarios, the performance is consistent across land cover types: the bias-corrected RF demonstrates close performance to the MDS and the RF performs worse than MDS for short gap length. For longer gap length, RF and bias-corrected RF significantly outperform the MDS. Considering that in the FLUXNET2015 dataset, long gaps account for 44% of the data, the bias-corrected RF can serve as a more reliable alternative algorithm to the MDS for hourly-scale data gap-filling, yielding more robust results than those produced by the MDS. Overall, the bias-corrected RF algorithm combines the superior performance of the original RF algorithm under long gap length scenario and provides corrections where the original RF underperforms.

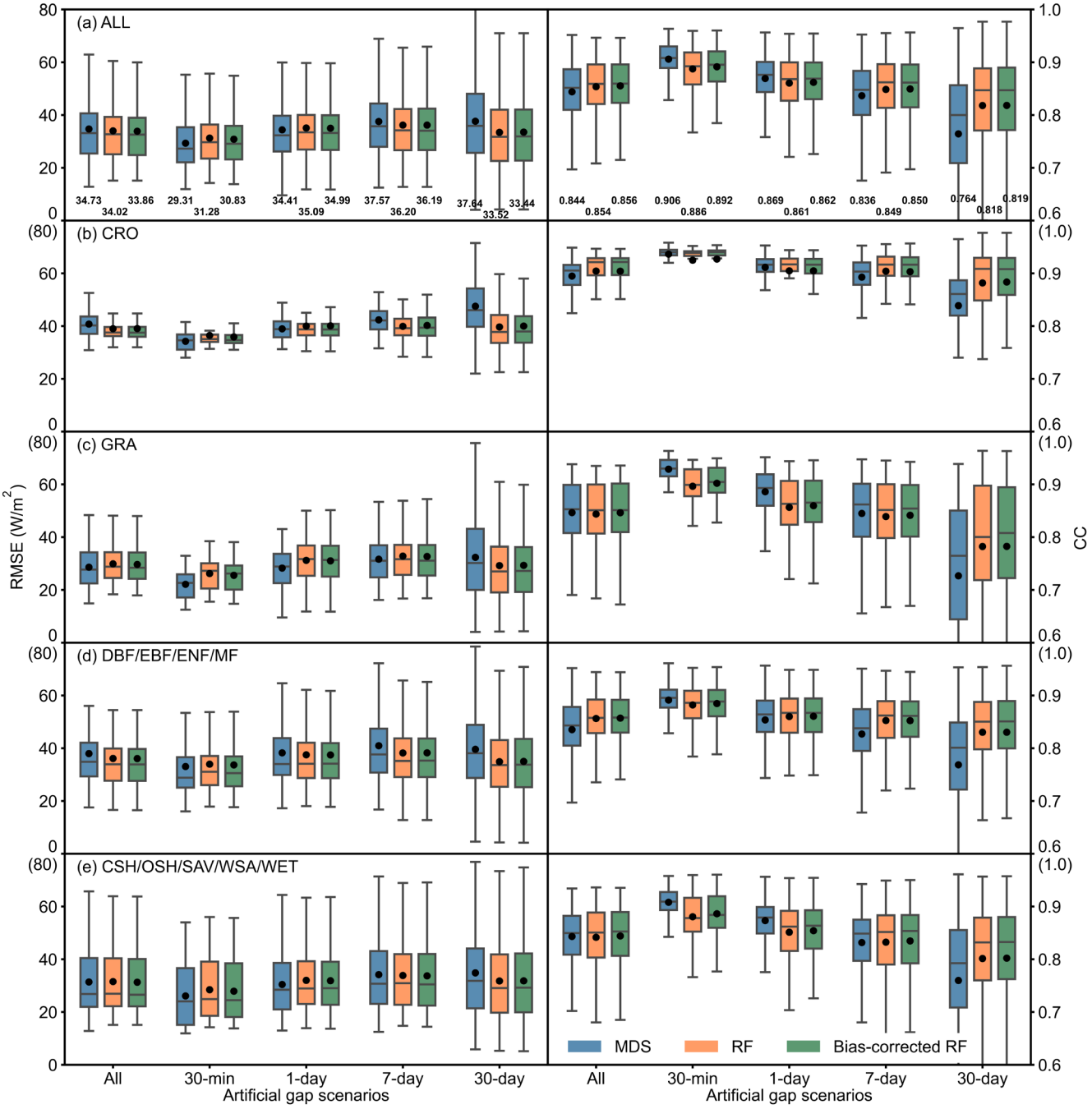
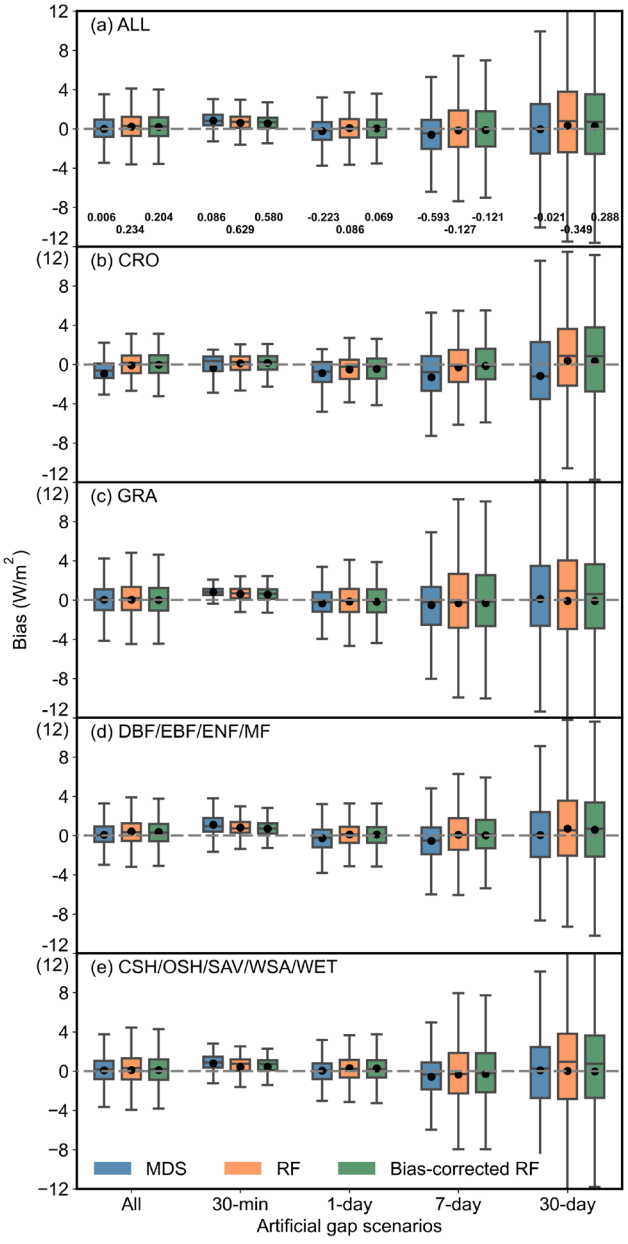


Figure 4 The gap-filling performance of three algorithms under different gap-length scenarios. The left panels show the results of the root mean square error (RMSE, W/m^2) and the right panels show results of correlation coefficient (CC) between gap-filled values and observations. Different rows of this figure indicate different land cover types. The three

horizontal lines of the boxes indicate the first quartile, median, and third quartile, respectively, and the black dots indicate the means. Data labels in this figure are the mean value of RMSE and CC. MDS: marginal distribution sampling. RF: random forest.



295 **Figure 5** The bias between gap-filled values and observations of three methods under different gap-length scenarios. Different rows of this figure indicate different land cover types. The three horizontal lines of the boxes indicate the first

quartile, median, and third quartile, respectively, and the black dots indicate the means. Data labels in this figure are the mean value of Bias. MDS: marginal distribution sampling. RF: random forest.

4.1.2 Examples of gap-filled data under artificial 30-day gap-length scenario

For the 30-day gap scenario, the bias-corrected RF algorithm performs better than MDS in characterizing time series. As illustrated in Figure 6, the bias-corrected RF demonstrates strong performance across all land cover types and provides a more accurate representation of daily periodic variations. Although minor biases persist in predicting certain extreme values, these are generally smaller compared to those produced by MDS. In contrast, MDS exhibits significant gap-filling biases across different land cover types, resulting in abnormal overestimations and underestimations (Fig 6a, b, and i). In some cases, it even fails to capture the daily variations of LE (Fig 6e), while also distorting irregular LE changes (Fig 6c).

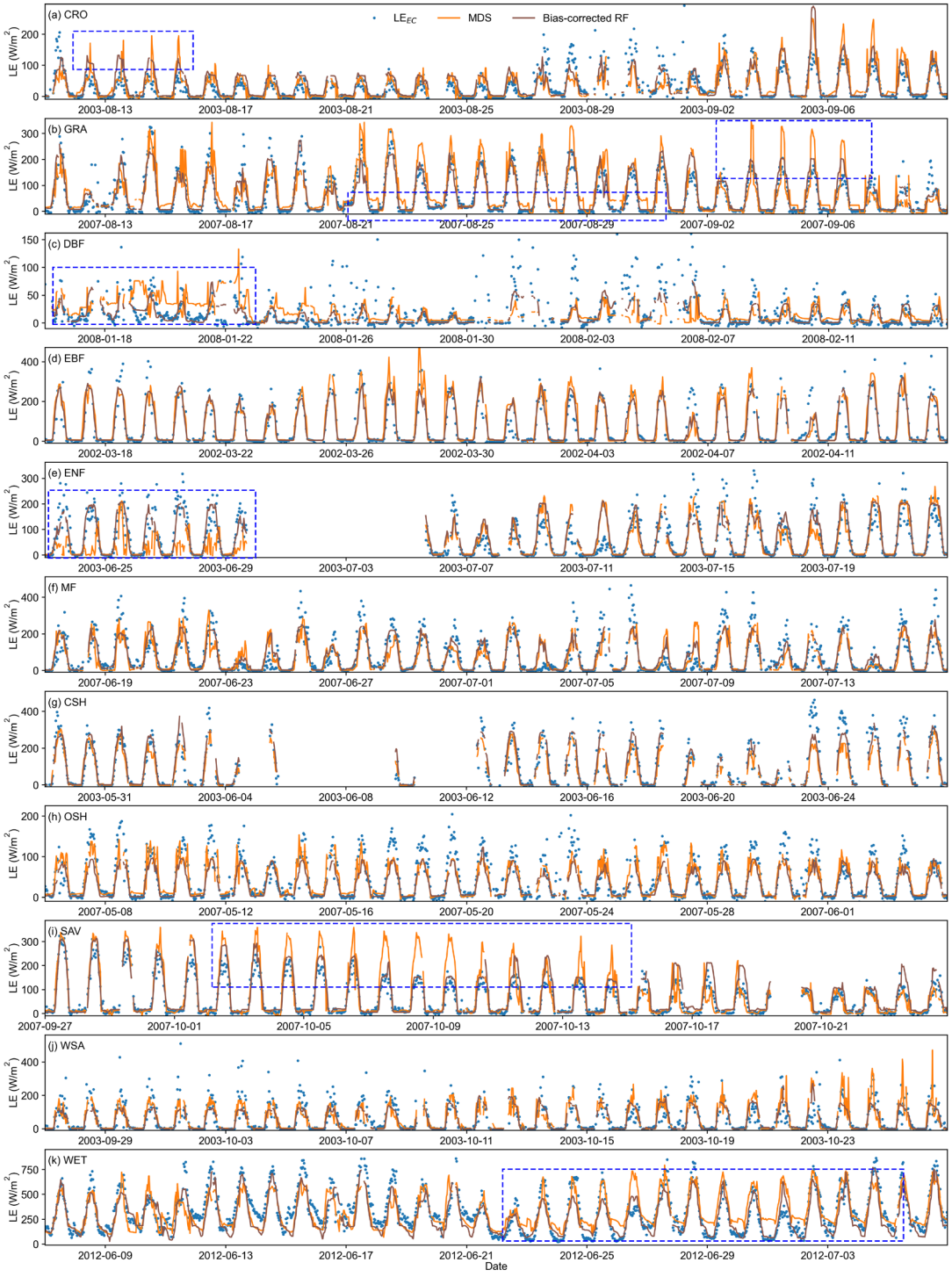


Figure 6 Time series of gap-filled results obtained from the bias-corrected RF algorithm compared to those from the MDS algorithm under artificial 30-day gap-length scenario across different land cover types. The blue dashed boxes indicate

scenarios where the MDS gap-filling results are significantly biased. The sites corresponding to each land cover type are: US-ARM, CN-Cng, FR-Fon, BR-Sa1, RU-Fyo, CA-Gro, US-KS2, ES-LJu, SD-Dem, AU-How, and US-Myb.

4.2 Evaluation of daily prolonged LE

4.2.1 The consistency between forward and backward prolongation.

As shown in Fig 7a and 7b, the prolongation performance in both forward and backward directions exhibit high consistency. The results have good accuracy, with RMSE (CC) of 16.58 W/m^2 (0.91) for forward and 17.35 W/m^2 (0.90) for backward. The slightly difference may be mainly due to a higher volume of missing data in the first two-thirds of the data compared to the last two-thirds for sites of these land cover types (See Section 5.1). There are slight variations in prolongation results for different land cover types (Fig 7c and 7d). Performance of CRO and DBF/EBF/ENF/MF is almost the same in both directions. Similar to the half-hourly data gap-filling, our results also demonstrate excellent performance in cropland, with a CC of 0.93 in both directions. GRA and CSH/OSH/SAV/WSA/WET perform slightly worse (2.46 W/m^2 and 3.74 W/m^2 higher) in the backward direction.

Figure 2b indicates that the need for forward prolongation is significantly greater than for backward prolongation from 2000 to 2022. Therefore, the validation in the following sections will focus only on the forward direction.

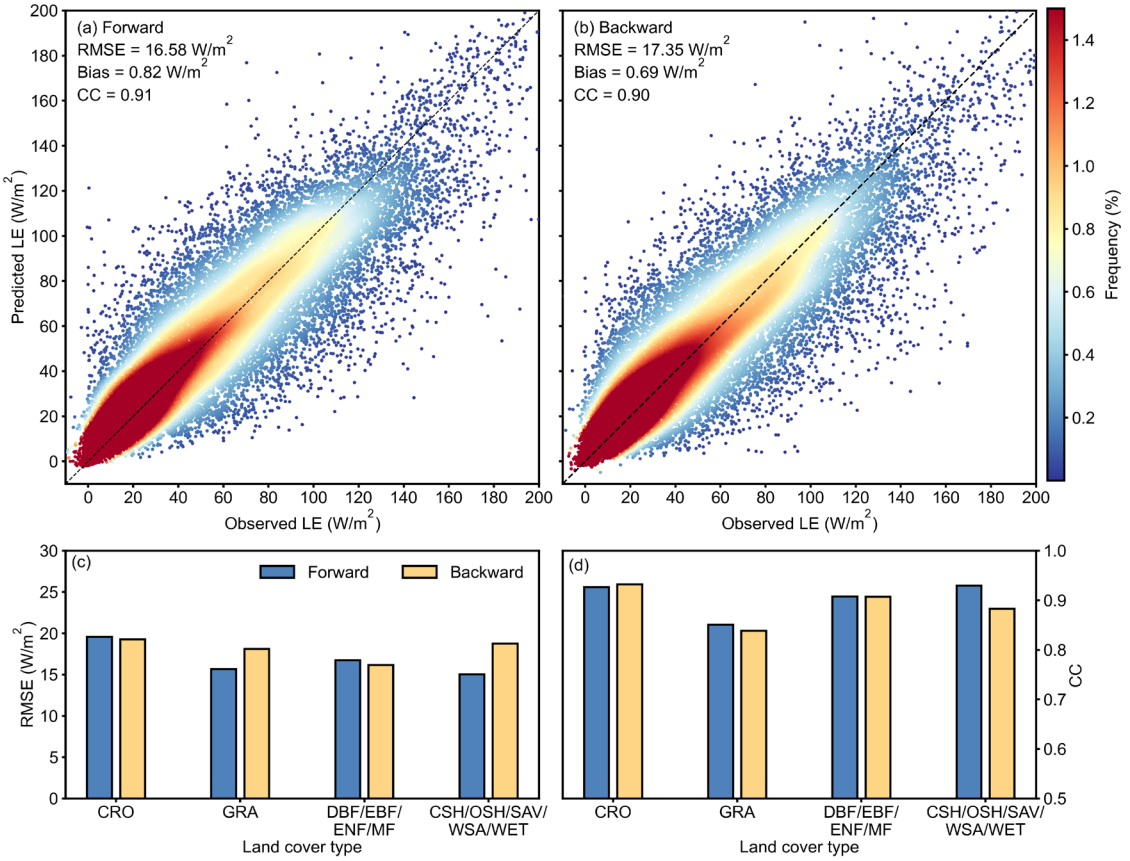


Figure 7 The consistency of forward and backward prolongation. (a) and (b) show the scatterplots of predicted daily LE against observations for forward and backward prolongation, respectively. (c) and (d) are the specific performance of different land cover types.

4.2.2 The temporal stability of the prolongation

We used data from the first three years and the first eight years for training and evaluated the prolongation performance for each subsequent year. Three years of data represents an extreme case of the minimum training data volume in this dataset, while eight years of data reflects a typical scenario within the dataset. Figure 8 shows that our prolongation results exhibit minimal performance degradation over time. The greater the amount of training data, the higher the temporal stability will be. Specifically, the model trained using the first three years yields CVs of RMSE and CC of only 3.29% and 3.83%, respectively. The model trained using the first eight years yields CVs of RMSE and CC of only 3.24% and 1.75%, respectively. The bias fluctuates within a small range around zero each year, indicating that our estimation bias is relatively robust. For different land cover types, DBF/EBF/ENF/MF shows good stability. GRA and CSH/OSH/SAV/WSA/WET show more noticeable fluctuations over time but did not experience significant performance degradation. Overall, our model demonstrates excellent temporal stability in both extreme and typical cases.

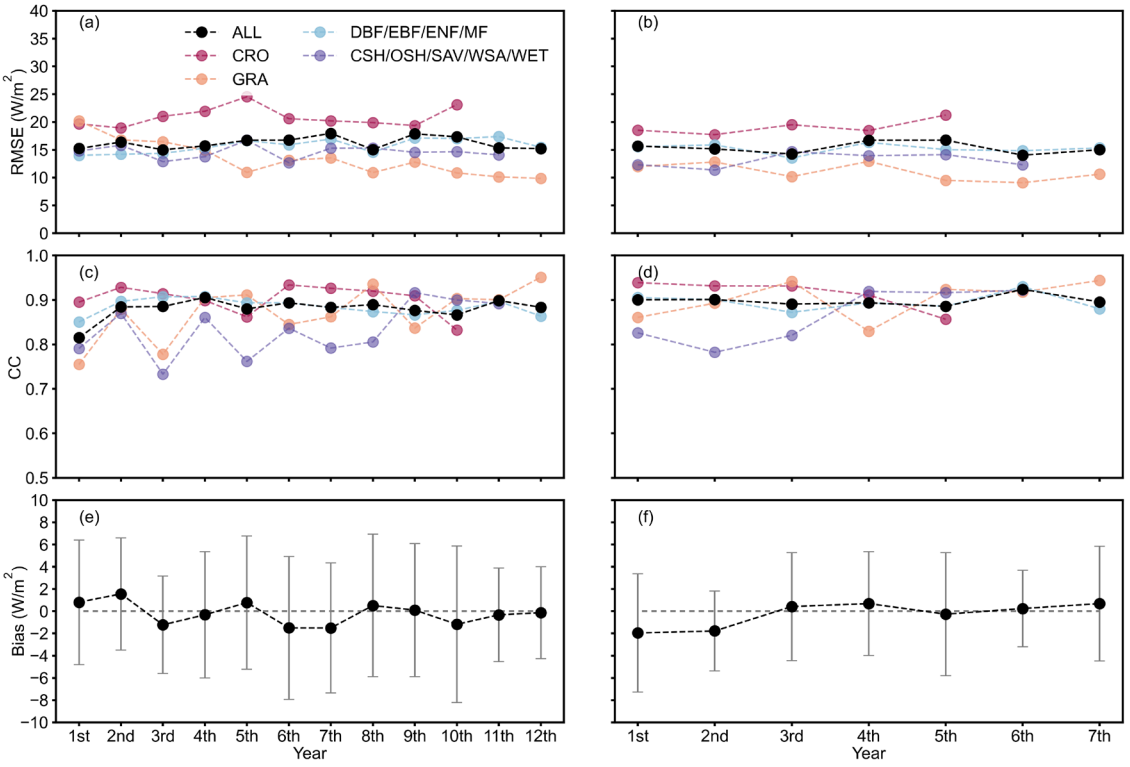


Figure 8 The temporal stability of the prolongation algorithm for different land cover types. (a), (c), and (e) show the median of RMSE, CC and Bias obtained from the model trained by first 3 years data, respectively. (b), (d), and (f) show the median of RMSE, CC and Bias of obtained from the model trained by first 8 years data, respectively.

4.2.3 Demonstration of daily- and monthly-scale prolonged time series

Due to the scarcity of days with a missing rate below 10%, we chose to compare the prolonged results from section 4.2.2 with the daily data aggregated from the hourly gap-filled data. We plotted the results obtained in section 4.2.2 as time series graphs and compared the prolonged results with the aggregated daily data from hourly gap-filled results. As shown in Fig 9 and 10, our prolongation algorithm effectively captures the seasonal variation of LE, aligning well with hourly gap-filled results in both magnitude and trend. The model performs excellently in both extreme (3 years data) and typical (8 years data) cases, particularly for sites with a land cover type of CRO. For evergreen vegetation sites (ENF and EBF) and sparse vegetation sites (SAV and OSH), the lack of vegetation change information leads to unclear influencing factors on LE variation. Some extreme high values are underestimated. However, our algorithm still performs well in capturing daily fluctuations.

Given that many global change studies focus on monthly scales, we aggregated both the daily data to assess the performance. As shown in Fig 11, the monthly scale results meet the requirements of related research. Both the trend and magnitude align well with hourly gap-filled results. The CRO sites match almost perfectly with the hourly gap-filled results, while the ENF and EBF sites, which performed slightly worse at the daily scale, accurately capture subtle fluctuations at the monthly scale.

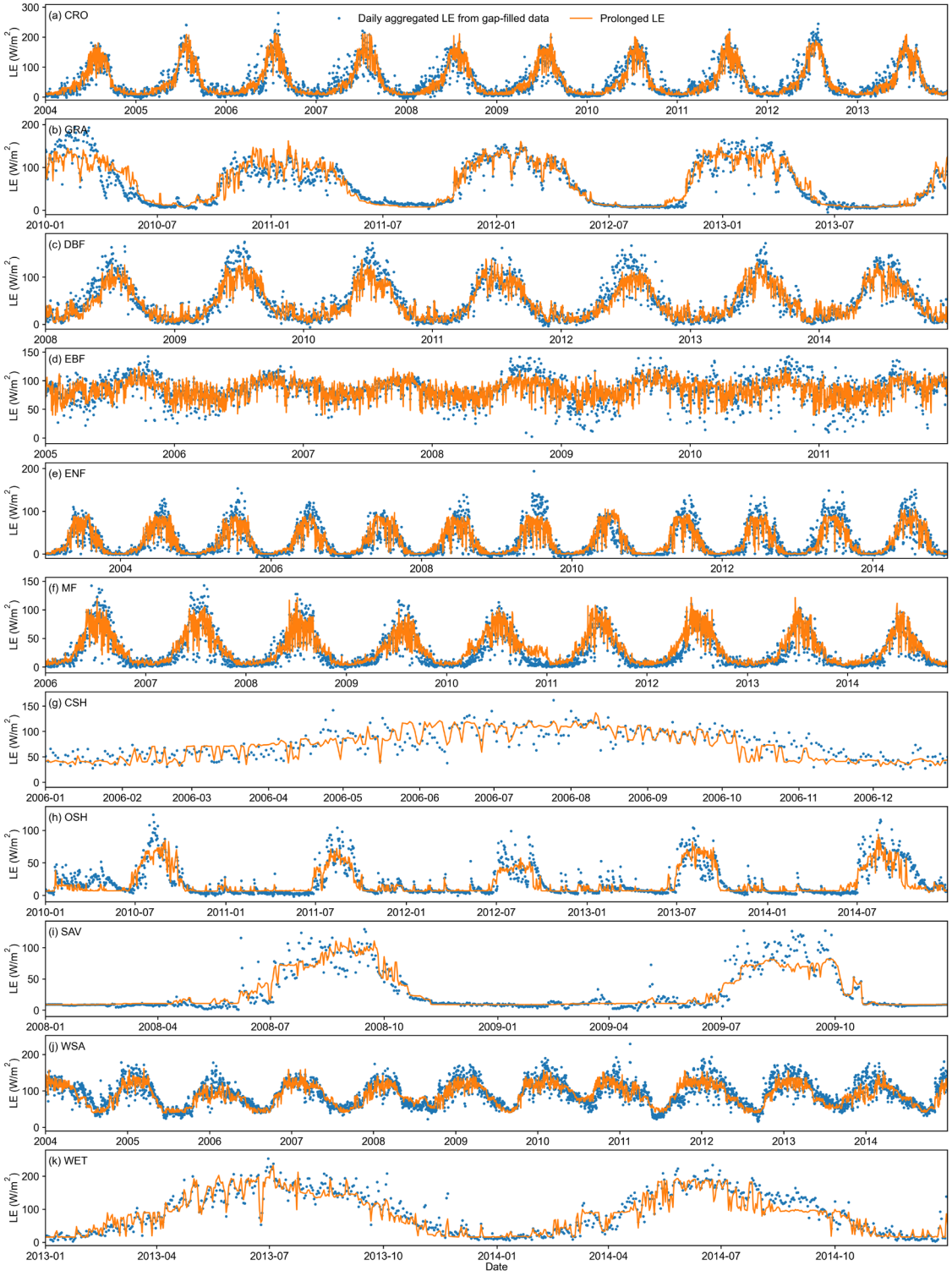

(a) CRO
Daily aggregated LE from gap-filled data
Prolonged LE
(b) GRA
(c) DBF
(d) EBF
(e) ENF
(f) MF
(g) CSH
(h) OSH
(i) SAV
(j) WSA
(k) WET
LE (W/m^2)
Date

Figure 9 Time series of daily prolonged results obtained from the model trained using the first three years across different land cover types. The sites corresponding to each land cover type are: US-Ne1, AU-DaP, FR-Fon, BR-Sa1, RU-Fyo, CA-Gro, US-KS2, US-Whs, SD-Dem, AU-How, and US-Myb.

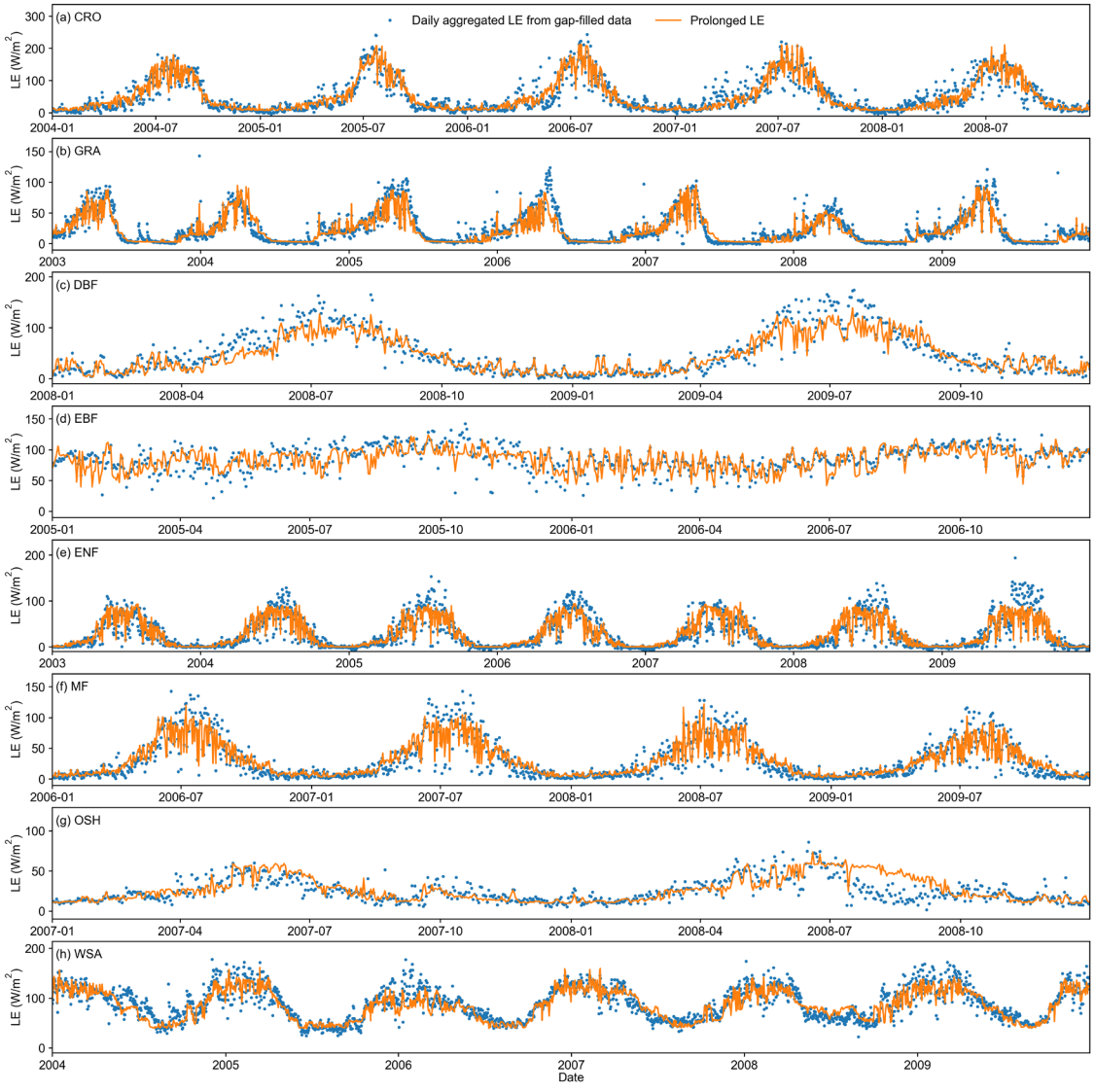
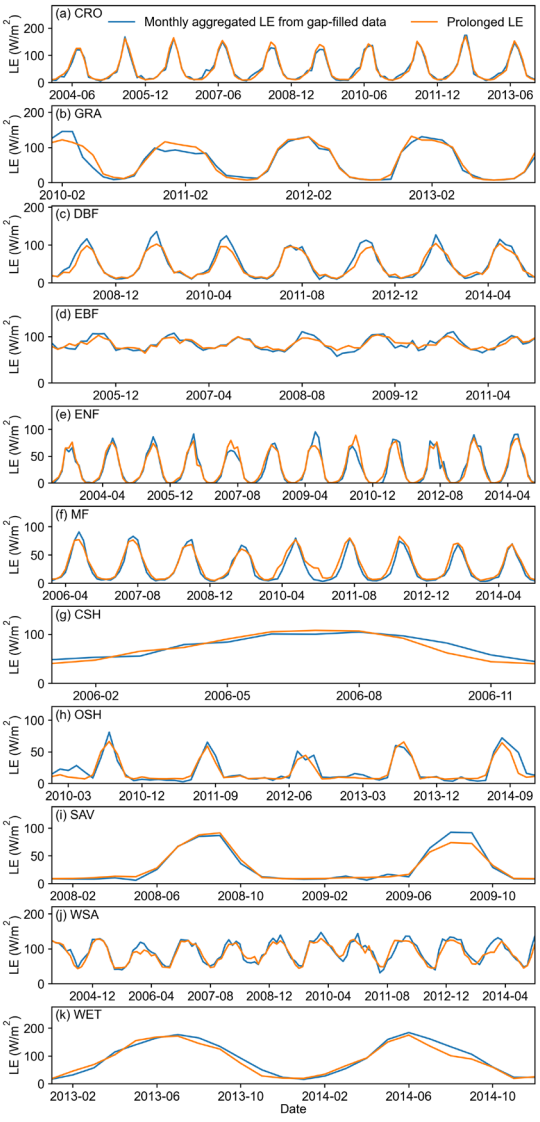


Figure 10 Time series of daily prolonged results obtained from the model trained using the first eight years across different land cover types. The sites corresponding to each land cover type are: US-Ne1, US-Var, FR-Fon, BR-Sa1, RU-Fyo, CA-Gro, ES-LJu, and AU-How.



365 **Figure 11** Time series of monthly aggregated results obtained from the model trained using the first three years across different land cover types. The stations corresponding to each land cover type are: US-Ne1, AU-DaP, FR-Fon, BR-Sa1, RU-Fyo, CA-Gro, US-KS2, US-Whs, SD-Dem, AU-How, and US-Myb.

5 Discussions

5.1 Comparison between FLUXNET2015 and our dataset

After extensive analysis of the experimental design results in Section 4, we have demonstrated excellent gap-filling and prolongation performance at the methodological level. To evaluate our released dataset, we compared it with the official dataset from FLUXNET2015. This is because missing data in observations cannot provide a verifiable truth. Figure 12 shows the data distribution results of gap-filled data at both hourly and daily scales for the two datasets. The results indicate a high consistency in data distribution between our dataset and FLUXNET2015. At the hourly scale, the median and quartiles of both datasets are nearly identical. For CRO, FLUXNET2015 exhibits slightly higher values compared to our dataset, while for GRA and CSH/OSH/SAV/WSA/WET, its estimates are slightly lower. At the daily scale, the consistency is even greater, with almost identical data distributions across all land surface types.

Additionally, we compared the differences between the two datasets aggregated to monthly and yearly scales. As shown in Fig 13, the data from both datasets distributes along the 1:1 line at both monthly and yearly scales. Although some months and years exhibited discrepancies between the two datasets, it still demonstrates a high degree of consistency. Specifically, at the monthly scale, we observed instances where some LE data of FLUXNET2015 show close values, while our predictions demonstrate clear distinctions. When aggregated to the yearly scale, these discrepancies manifested as outliers. This instance arises because many FLUXNET2015 sites experienced complete data loss for the first four to eight months (e.g., AU-ASM from January to August 2010, CA-Gro from January to July 2003, US-UMd from January to April 2007, among others). Due to the lack of neighboring information in the sliding window, the MDS algorithm struggled to provide effective gap-filling, resulting in nearly identical gap-filled values for those months. Consequently, these months could not be included in the usable data range, rendering the aggregated results at the yearly scale unreliable. In contrast, our algorithm can utilize the reference data for each specific moment to predict the corresponding LE, so we can provide more accurate gap-filling results.

Therefore, the advantages of our dataset are: 1) Hourly scale gap-filling enhances accuracy compared to FLUXNET2015 under long gap-length scenarios; 2) daily scale results show good consistency with FLUXNET2015 while providing a much longer time series (23 years compared to averaged 8 years). However, our data does have some limitations. For instance, due to the restrictions of NDVI data, our dataset only provides data from February 18, 2000 for both hourly gap-filling and daily prolongation.

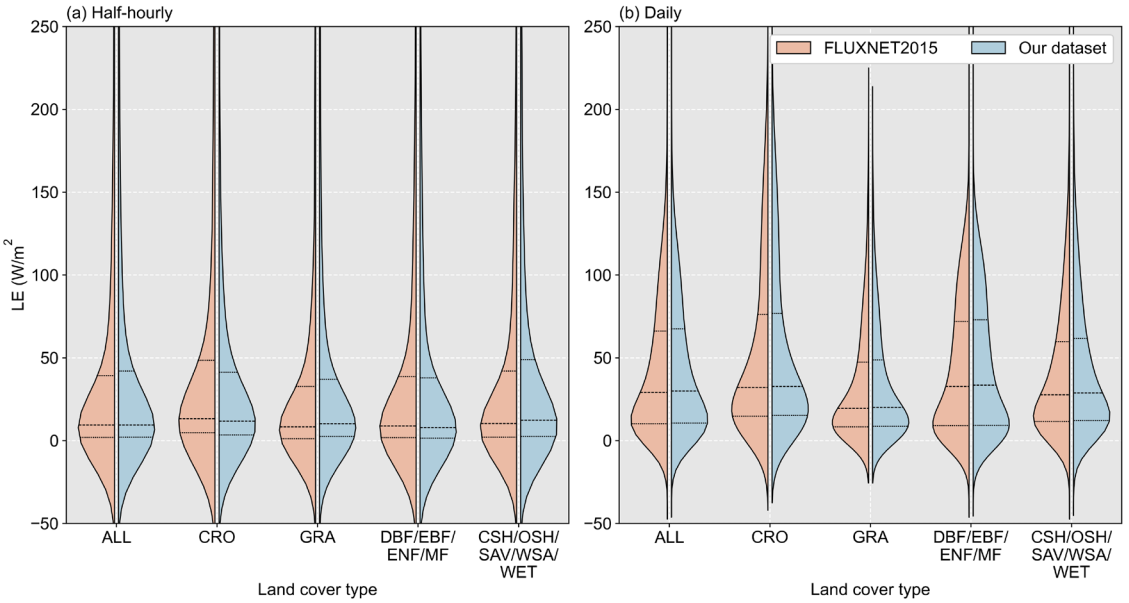


Figure 12 Distribution of gap-filled data at both (a) half-hourly and (b) daily scales for our dataset and FLUXNET2015 dataset.

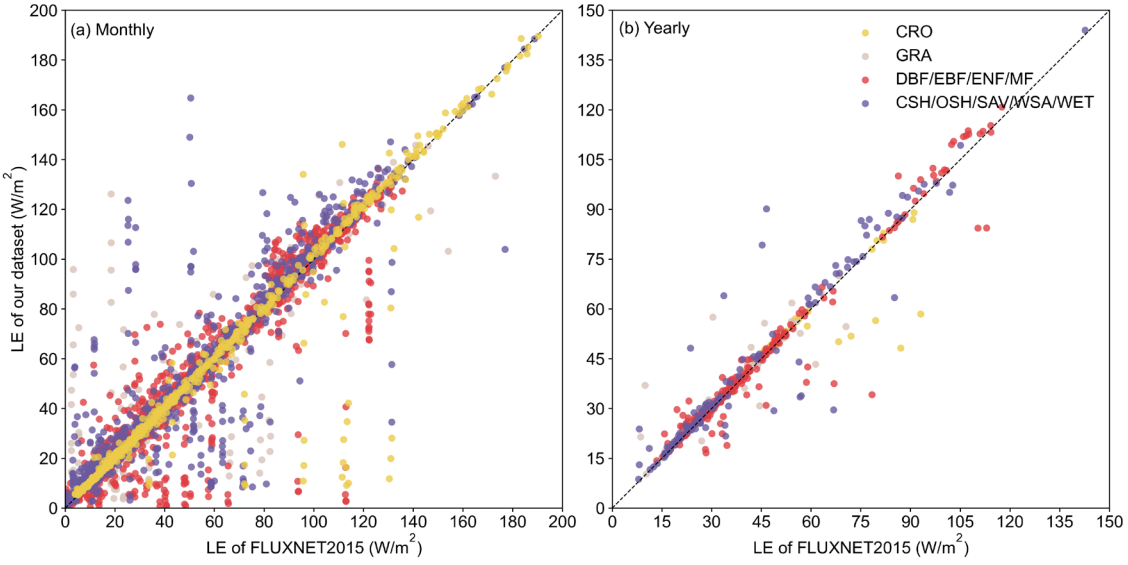


Figure 13 Scatterplot of LE data of our dataset against that of FLUXNET2015 dataset.

5.2 Reference variables importance analysis

Figure 14 presents the results of the reference variables importance using the permutation feature importance technique. Each input feature is randomly shuffled to calculate the performance deterioration. For half-hourly or hourly gap-filling, the

order of variable importance is SW_IN > NDVI > TA > LW_IN > RH > WS > PA. Consistent with earlier research (Irvin et al., 2021; Zhu et al., 2022; Li et al., 2024a), SW_IN is the key variable that significantly influences LE variations across terrestrial ecosystems. It provides energy for the ET process. Throughout the day, SW_IN exhibits significant diurnal variation. NDVI is the second most important variable, but its influence varies between sites. This explains why the performance of the two land cover types in section 4.2.3 is slightly inferior to that of other types. For sites with evergreen vegetation, seasonal changes in vegetation are not pronounced, making NDVI less effective in providing clear information to the model. For daily prolongation, the order of variable importance is different. The importance of SW_IN decreases significantly because daily LE variation is more closely related to NDVI, which reflects seasonal changes. Similar to the hourly scale, NDVI also shows inconsistencies between sites for the same reasons. Additionally, TA, as the third most important variable, provides critical information at sites dominated by soil evaporation. Variables like LW_IN, RH, WS, and PA hold comparable significance as minor factors, offering insights into the meteorological background conditions.

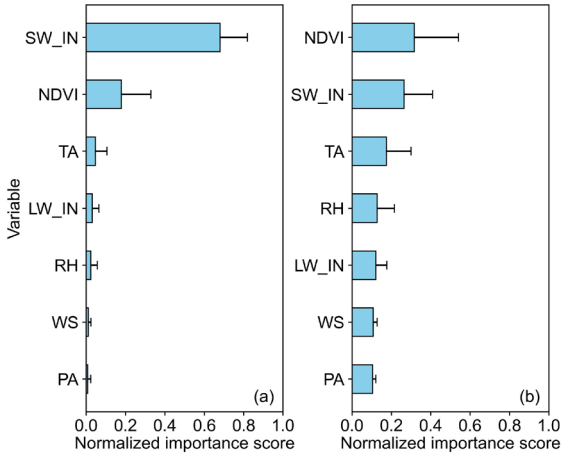


Figure 14 Variables importance for half-hourly or hourly gap-filling and daily prolongation.

5.3 Advantages and disadvantages

Our study presents several notable advantages: 1) The bias-corrected RF shows better performance than the official MDS approach, especially for filling very long gaps (up to 30 days). Additionally, it allows for temporal prolongation, which the MDS method cannot achieve. Furthermore, our method enables the incorporation of a broader range of reference variables to establish a more robust non-linear relationship between LE and its drivers; 2) Compared to the FLUXNET2015 dataset, our hourly gap-filled data show improved quality and simpler implementation. The daily prolonged data provide extended temporal coverage (2000-2022) that is particularly valuable for evapotranspiration (ET) modeling and global-scale studies. However, some limitations in terms of variable importance, sensitivity and stability merit further discussion. The variable importance analysis (Section 5.2) indicated our method exhibits strong sensitivity to SW_IN data for gap-filling and to

NDVI for prolongation. While we implemented bias correction between ground observations and ERA5-Land data, potential quality issues in SW_IN and NDVI inputs may still affect final results. Future improvements could incorporate higher-quality input data with more stable biases to enhance result reliability.

6 Data availability

Our released dataset mainly contains four types of data:

(1) Half-hourly or hourly gap-filled data: The data were well gap-filled data using the novel bias-corrected RF algorithm. Filenames include "HH" for half-hourly or "HR" for hourly data. The time format follows FLUXNET2015 standards, with paired timestamps recorded in local time. The start and end times align with the observation period at each site. For QC flags, a value of 0 indicates observed data, while 1 indicates gap-filled data.

(2) Aggregated daily data. This daily dataset was aggregated from the gap-filled half-hourly data to a daily scale. The start and the end times match the observation period at each site. QC flags represent the percentage of valid hourly observations for each day.

(3) Prolonged daily data: This dataset provides the prolonged daily data using the novel bias-corrected RF algorithm. The seamless data spans 2000-02-18 to 2022-12-31. For the prolonged part, the quality flag is set to 2. The rest is consistent with the aggregated daily data.

(4) Aggregated monthly and yearly data: These datasets were aggregated from the prolonged daily data. QC flags indicate the proportion of days with >90% valid hourly observations per month or year. No distinction is made between prolonged data and completely missing daily data. The time span for the monthly data is 2000-03 to 2022-12, and that for the yearly data is 2001-2022.

All files are formatted as csv files. NDVI and debiased reference variables from ERA5-Land are also provided in our released data. The product has been deposited at https://doi.org/10.5281/zenodo.13853409 (Li et al., 2024b) and can be downloaded publicly.

7 Conclusions

The current LE_{EC} data are increasingly insufficient to meet the growing demand for long time-series benchmark data to support climate change studies, model development, and product validation. To address these limitations in FLUXNET2015, we developed a gap-filling and prolongation framework for LE_{EC} data and established a benchmark dataset for ground-based ET (2000-2022) across 64 global sites. The results indicate that:

1) Hourly gap-filling: the novel bias-corrected RF algorithm demonstrates excellent performance, achieving a mean RMSE of 33.86 W/m^2. It improves the original RF algorithm's poor performance for short gaps, approaching the performance of official algorithm (MDS). For long gaps, it significantly outperforms the MDS algorithm by 11.15%. The algorithm more

accurately predicts extreme values, thereby reducing result uncertainty compared to MDS. It performs consistently well across various land surface types, with the most notable improvements observed in cropland. Additionally, our gap-filled data distribution shows strong agreement with official products.

2) Daily prolongation: our method exhibits robust performance in both forward and backward directions (16.58 W/m^2 and 17.35 W/m^2, respectively). The method shows slight variations in performance across different land surface types, with the best performance for cropland. In terms of temporal stability, our results maintain excellent performance under both extreme conditions (training with the first three years of data) and typical conditions (training with the first eight years of data). The time series effectively captures seasonal variations in LE, aligning well with observations.

3) For hourly data gap-filling, SW_IN is the most important factor, while NDVI plays a decisive role in daily prolongation. In cases where the land surface is dominated by evergreen or sparse vegetation, the importance of NDVI significantly decreases.

Overall, our proposed gap-filling and prolongation framework for LE_{EC} data is robust and a benchmark dataset for global ET estimation based on FLUXNET2015 from 2000 to 2022 is established. It can provide essential data support for ET modelling, water-carbon cycle studies, and climate impact assessments.

Appendix A: Site information

Site	IGBP	Latitude	Longitude	Start year	End year	Time cover after 2000	Hourly missing ratio
AU-ASM	SAV	-22.283	133.249	2010	2014	5	0.37
AU-Cpr	SAV	-34.0021	140.5891	2010	2014	5	0.28
AU-DaP	GRA	-14.0633	131.3181	2007	2013	7	0.36
AU-DaS	SAV	-14.1593	131.3881	2008	2014	7	0.21
AU-Dry	SAV	-15.2588	132.3706	2008	2014	7	0.45
AU-Gin	WSA	-31.3764	115.7138	2011	2014	4	0.44
AU-How	WSA	-12.4943	131.1523	2001	2014	14	0.35
AU-Rig	GRA	-36.6499	145.5759	2011	2014	4	0.26
AU-Stp	GRA	-17.1507	133.3502	2008	2014	7	0.29
AU-TTE	GRA	-22.287	133.64	2012	2014	3	0.40
AU-Whr	EBF	-36.6732	145.0294	2011	2014	4	0.32
AU-Wom	EBF	-37.4222	144.0944	2010	2014	5	0.41
BR-Sa1	EBF	-2.8567	-54.9589	2002	2011	10	0.27
BR-Sa3	EBF	-3.018	-54.9714	2000	2004	5	0.47
CA-Gro	MF	48.2167	-82.1556	2003	2014	12	0.24
CA-NS2	ENF	55.9058	-98.5247	2001	2005	5	0.49
CA-NS3	ENF	55.9117	-98.3822	2001	2005	5	0.33
CA-Oas	DBF	53.6289	-106.1978	1996	2010	11	0.17
CA-Qfo	ENF	49.6925	-74.3421	2003	2010	8	0.23

CA-SF1	ENF	54.485	-105.8176	2003	2006	4	0.37
CA-SF2	ENF	54.2539	-105.8775	2001	2005	5	0.35
CA-SF3	OSH	54.0916	-106.0053	2001	2006	6	0.36
CA-TP1	ENF	42.6609	-80.5595	2002	2014	13	0.47
CA-TP3	ENF	42.7068	-80.3483	2002	2014	13	0.41
CA-TP4	ENF	42.7102	-80.3574	2002	2014	13	0.18
CG-Tch	SAV	-4.2892	11.6564	2006	2009	4	0.57
CN-Cha	MF	42.4025	128.0958	2003	2005	3	0.24
CN-Cng	GRA	44.5934	123.5092	2007	2010	4	0.27
CN-Din	EBF	23.1733	112.5361	2003	2005	3	0.31
CN-Ha2	WET	37.6086	101.3269	2003	2005	3	0.17
CN-Qia	ENF	26.7414	115.0581	2003	2005	3	0.21
DE-Obe	ENF	50.7867	13.7213	2008	2014	7	0.18
DE-Tha	ENF	50.9626	13.5651	1996	2014	15	0.13
ES-Amo	OSH	36.8336	-2.2523	2007	2012	6	0.38
ES-LJu	OSH	36.9266	-2.7521	2004	2013	10	0.27
FR-Fon	DBF	48.4764	2.7801	2005	2014	10	0.18
GF-Guy	EBF	5.2788	-52.9249	2004	2014	11	0.24
MY-PSO	EBF	2.973	102.3062	2003	2009	7	0.21
RU-Fyo	ENF	56.4615	32.9221	1998	2014	15	0.22
SD-Dem	SAV	13.2829	30.4783	2005	2009	5	0.64
US-AR1	GRA	36.4267	-99.42	2009	2012	4	0.23
US-AR2	GRA	36.6358	-99.5975	2009	2012	4	0.33
US-ARM	CRO	36.6058	-97.4888	2003	2012	10	0.15
US-Blo	ENF	38.8953	-120.6328	1997	2007	8	0.36
US-Goo	GRA	34.2547	-89.8735	2002	2006	5	0.40
US-KS2	CSH	28.6086	-80.6715	2003	2006	4	0.27
US-Me2	ENF	44.4526	-121.5589	2002	2014	13	0.14
US-Me3	ENF	44.3154	-121.6078	2004	2009	6	0.26
US-MMS	DBF	39.3232	-86.4131	1999	2014	15	0.34
US-Myb	WET	38.0499	-121.765	2010	2014	5	0.32
US-Ne1	CRO	41.1651	-96.4766	2001	2013	13	0.15
US-Ne2	CRO	41.1649	-96.4701	2001	2013	13	0.23
US-Ne3	CRO	41.1797	-96.4397	2001	2013	13	0.21
US-NR1	ENF	40.0329	-105.5464	1998	2014	15	0.22
US-SRC	OSH	31.9083	-110.8395	2008	2014	7	0.38
US-SRG	GRA	31.7894	-110.8277	2008	2014	7	0.14
US-SRM	WSA	31.8214	-110.8661	2004	2014	11	0.12
US-Ton	WSA	38.4309	-120.966	2001	2014	14	0.31
US-Twt	CRO	38.1087	-121.6531	2009	2014	6	0.36

US-UMB	DBF	45.5598	-84.7138	2000	2014	15	0.23
US-UMd	DBF	45.5625	-84.6975	2007	2014	8	0.17
US-Var	GRA	38.4133	-120.9508	2000	2014	15	0.21
US-Whs	OSH	31.7438	-110.0522	2007	2014	8	0.16
US-Wkg	GRA	31.7365	-109.9419	2004	2014	11	0.16

Author contributions

Conceptualization: WL and YC. Methodology: WL, ZY, and YC. Data curation: WL. Funding acquisition: YC. Writing (initial): WL. Writing (review and editing): ZY, YQ, HY, LS, LW, YS, and YC. Supervision: YC.

Competing interests

The contact author has declared that none of the authors has any competing interests.

Acknowledgements

The authors would like to thank the scikit-learn (https://scikit-learn.org/stable/install.html) team and the ReddyProc (https://cran.r-project.org/web/packages/REddyProc/index.html) team for the packages that help their method establishment and validation. They also thank the FLUXNET and the research groups for providing the CC-BY-4.0 (Tier one) open-access eddy covariance data (https://fluxnet.org/data/fluxnet2015-dataset/). They thank the ECWMF team for the public ERA5-Land reanalysis data (https://www.ecmwf.int/en/era5-land) and the MODIS science team for the MYD13Q1 data. Additionally, they also thank the Google Earth Engine platform for downloading ERA5-Land and MYD13Q1 data efficiently.

Financial support

This study was financially supported by the National Natural Science Foundation of China (Grant No. 42471375 and No. 42130104) and the Key R&D Program of the Ministry of Science and Technology, China (Grant No. 2022YFC3002802).

References

Aubinet, M., Vesala, T., and Papale, D.: Eddy covariance: A practical guide to measurement and data analysis, Springer Science & Business Media, https://doi.org/10.1007/978-94-007-2351-1, 2012.

Baldocchi, D. D.: How eddy covariance flux measurements have contributed to our understanding of global change biology, Glob. Change Biol., 26, 242-260, https://doi.org/10.1111/gcb.14807, 2020.

Breiman, L.: Random forests, Mach. Learn., 45, 5-32, https://doi.org/10.1023/A:1010933404324, 2001.

Chen, Y., Xia, J., Liang, S., Feng, J., Fisher, J. B., Li, X., Li, X., Liu, S., Ma, Z., and Miyata, A.: Comparison of satellite-based evapotranspiration models over terrestrial ecosystems in china, Remote Sens. Environ., 140, 279-293, https://doi.org/10.1016/j.rse.2013.08.045, 2014.

Cui, N., He, Z., Jiang, S., Wang, M., Yu, X., Zhao, L., Qiu, R., Gong, D., Wang, Y., and Feng, Y.: Inter-comparison of the penman-monteith type model in modeling the evapotranspiration and its components in an orchard plantation of southwest china, Agric. Water Manage., 289, 108541, https://doi.org/10.1016/j.agwat.2023.108541, 2023.

Cui, Y. and Jia, L.: Estimation of evapotranspiration of "soil-vegetation" system with a scheme combining a dual-source model and satellite data assimilation, J. Hydrol., 603, 127145, https://doi.org/10.1016/j.jhydrol.2021.127145, 2021.

Cui, Y., Jia, L., and Fan, W.: Estimation of actual evapotranspiration and its components in an irrigated area by integrating the shuttleworth-wallace and surface temperature-vegetation index schemes using the particle swarm optimization algorithm, Agric. For. Meteorol., 307, 108488, https://doi.org/10.1016/j.agrformet.2021.108488, 2021a.

Cui, Y., Song, L., and Fan, W.: Generation of spatio-temporally continuous evapotranspiration and its components by coupling a two-source energy balance model and a deep neural network over the heihe river basin, J. Hydrol., 597, 126176, https://doi.org/10.1016/j.jhydrol.2021.126176, 2021b.

Das, N. N., Entekhabi, D., Dunbar, R. S., Colliander, A., Chen, F., Crow, W., Jackson, T. J., Berg, A., Bosch, D. D., and Caldwell, T.: The smap mission combined active-passive soil moisture product at 9 km and 3 km spatial resolutions, Remote Sens. Environ., 211, 204-217, https://doi.org/10.1016/j.rse.2018.04.011, 2018.

Feng, P., Wang, B., Liu, D. L., and Yu, Q.: Machine learning-based integration of remotely-sensed drought factors can improve the estimation of agricultural drought in south-eastern australia, Agric. Syst., 173, 303-316, https://doi.org/10.1016/j.agsy.2019.03.015, 2019.

Foltýnová, L., Fischer, M., and McGloin, R. P.: Recommendations for gap-filling eddy covariance latent heat flux measurements using marginal distribution sampling, Theor. Appl. Climatol., 139, 677-688, https://doi.org/10.1007/s00704-019-02975-w, 2020.

Hu, X., Shi, L., Lin, G., and Lin, L.: Comparison of physical-based, data-driven and hybrid modeling approaches for evapotranspiration estimation, J. Hydrol., 601, 126592, https://doi.org/10.1016/j.jhydrol.2021.126592, 2021.

Irvin, J., Zhou, S., McNicol, G., Lu, F., Liu, V., Fluet-Chouinard, E., Ouyang, Z., Knox, S. H., Lucas-Moffat, A., and Trotta, C.: Gap-filling eddy covariance methane fluxes: Comparison of machine learning model predictions and uncertainties at fluxnet-ch4 wetlands, Agric. For. Meteorol., 308, 108528, https://doi.org/10.1016/j.agrformet.2021.108528, 2021.

Li, W., Yao, Z., Pan, X., Wei, Z., Jiang, B., Wang, J., Xu, M., and Cui, Y.: A ground-independent method for obtaining complete time series of in situ evapotranspiration observations, J. Hydrol., 632, 130888, https://doi.org/10.1016/j.jhydrol.2024.130888, 2024a.

Li, W., Yao, Z., Qu, Y., Yang, H., Song, Y., Song, L., Wu, L., and Cui, Y.: A benchmark dataset for global evapotranspiration estimation based on fluxnet2015 from 2000 to 2022 (v1.0) [dataset], https://doi.org/10.5281/zenodo.13853409, 2024b.

Mahabbati, A., Beringer, J., Leopold, M., McHugh, I., Cleverly, J., Isaac, P., and Izady, A.: A comparison of gap-filling algorithms for eddy covariance fluxes and their drivers, Geosci. Instrum. Methods Data Syst., 10, 123-140, https://doi.org/10.5194/gi-10-123-2021, 2021.

Martens, B., Miralles, D. G., Lievens, H., Van Der Schalie, R., De Jeu, R. A., Fernández-Prieto, D., Beck, H. E., Dorigo, W. A., and Verhoest, N. E.: Gleam v3: Satellite-based land evaporation and root-zone soil moisture, Geosci. Model Dev., 10, 1903-1925, https://doi.org/10.5194/gmd-10-1903-2017, 2017.

Miralles, D. G., Bonte, O., Koppa, A., Baez-Villanueva, O. M., Tronquo, E., Zhong, F., Beck, H. E., Hulsman, P., Dorigo, W., Verhoest, N. E. C., and Haghdoost, S.: Gleam4: Global land evaporation and soil moisture dataset at 0.1° resolution from 1980 to near present, Sci. Data, 12, 10.1038/s41597-025-04610-y, 2025.

Moffat, A. M., Papale, D., Reichstein, M., Hollinger, D. Y., Richardson, A. D., Barr, A. G., Beckstein, C., Braswell, B. H., Churkina, G., and Desai, A. R.: Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes, Agric. For. Meteorol., 147, 209-232, https://doi.org/10.1016/j.agrformet.2007.08.011, 2007.

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

Mu, Q., Zhao, M., and Running, S. W.: Improvements to a modis global terrestrial evapotranspiration algorithm, Remote Sens. Environ., 115, 1781-1800, https://doi.org/10.1016/j.rse.2011.02.019, 2011.

Oki, T. and Kanae, S.: Global hydrological cycles and world water resources, Science, 313, 1068-1072, https://10.1126/science.1128845, 2006.

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. d., Ligne, A. D., De Oliveira, R. C., Delpierre,

N., Desai, A. R., Di Bella, C. M., Tommasi, P. d., 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., Gharun, M., Gianelle, D., Gielen, B., Gioli, B., Gitelson, A., Goded, I., Goeckede, M., Goldstein, A. H., Gough, C. M., Goulden, M. L., Graf, A., Griebel, A., Gruening, C., Grünwald, T., Hammerle, A., Han, S., Han, X., Hansen, B. U., Hanson, C., Hatakka, J., He, Y., Hehn, M., Heinesch, B., Hinko-Najera, N., Hörtnagl, L., Hutley, L., Ibrom, A., Ikawa, H., Jackowicz-Korczynski, M., Janouš, D., Jans, W., Jassal, R., Jiang, S., Kato, T., Khomik, M., Klatt, J., Knohl, A., Knox, S., Kobayashi, H., Koerber, G., Kolle, O., Kosugi, Y., Kotani, A., Kowalski, A., Kruijt, B., Kurbatova, J., Kutsch, W. L., Kwon, H., Launiainen, S., Laurila, T., Law, B., Leuning, R., Li, Y., Liddell, M., Limousin, J.-M., Lion, M., Liska, A. J., Lohila, A., López-Ballesteros, A., López-Blanco, E., Loubet, B., Loustau, D., Lucas-Moffat, A., Lüers, J., Ma, S., Macfarlane, C., Magliulo, V., Maier, R., Mammarella, I., Manca, G., Marcolla, B., Margolis, H. A., Marras, S., Massman, W., Mastepanov, M., Matamala, R., Matthes, J. H., Mazzenga, F., McCaughey, H., McHugh, I., McMillan, A. M. S., Merbold, L., Meyer, W., Meyers, T., Miller, S. D., Minerbi, S., Moderow, U., Monson, R. K., Montagnani, L., Moore, C. E., Moors, E., Moreaux, V., Moureaux, C., Munger, J. W., Nakai, T., Neirynck, J., Nesic, Z., Nicolini, G., Noormets, A., Northwood, M., Nosetto, M., Nouvellon, Y., Novick, K., Oechel, W., Olesen, J. E., Ourcival, J.-M., Papuga, S. A., Parmentier, F.-J., Paul-Limoges, E., Pavelka, M., Peichl, M., Pendall, E., Phillips, R. P., Pilegaard, K., Pirk, N., Posse, G., Powell, T., Prasse, H., Prober, S. M., Rambal, S., Rannik, Ü., Raz-Yaseef, N., Rebmann, C., Reed, D., Dios, V. R. d., Restrepo-Coupe, N., Reverter, B. R., Roland, M., Sabbatini, S., Sachs, T., Saleska, S. R., Sánchez-Cañete, E. P., Sanchez-Mejia, Z. M., Schmid, H. P., Schmidt, M., Schneider, K., Schrader, F., Schroder, I., Scott, R. L., Sedlák, P., Serrano-Ortíz, P., Shao, C., Shi, P., Shironya, I., Siebicke, L., Šigut, L., Silberstein, R., Sirca, C., Spano, D., Steinbrecher, R., Stevens, R. M., Sturtevant, C., Suyker, A., Tagesson, T., Takanashi, S., Tang, Y., Tapper, N., Thom, J., Tomassucci, M., Tuovinen, J.-P., Urbanski, S., Valentini, R., van der Molen, M., van Gorsel, E., van Huissteden, K., Varlagin, A., Verfaillie, J., Vesala, T., Vincke, C., Vitale, D., Vygodskaya, N., Walker, J. P., Walter-Shea, E., Wang, H., Weber, R., Westermann, S., Wille, C., Wofsy, S., Wohlfahrt, G., Wolf, S., Woodgate, W., Li, Y., Zampedri, R., Zhang, J., Zhou, G., Zona, D., Agarwal, D., Biraud, S., Torn, M., and Papale, D.: 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.

Reichstein, M., Falge, E., Baldocchi, D., Papale, D., Aubinet, M., Berbigier, P., Bernhofer, C., Buchmann, N., Gilmanov, T., and Granier, A.: On the separation of net ecosystem exchange into assimilation and ecosystem respiration: Review and improved algorithm, Glob. Change Biol., 11, 1424-1439, https://doi.org/10.1111/j.1365-2486.2005.001002.x, 2005.

Song, Y., Guo, Y., Li, S., Li, W., and Jin, X.: Elevated co2 concentrations contribute to a closer relationship between vegetation growth and water availability in the northern hemisphere mid-latitudes, Environ. Res. Lett., 19, 084013, 10.1088/1748-9326/ad5f43, 2024.

Tang, R., Peng, Z., Liu, M., Li, Z.-L., Jiang, Y., Hu, Y., Huang, L., Wang, Y., Wang, J., Jia, L., Zheng, C., Zhang, Y., Zhang, K., Yao, Y., Chen, X., Xiong, Y., Zeng, Z., and Fisher, J. B.: Spatial-temporal patterns of land surface evapotranspiration from global products, Remote Sens. Environ., 304, 114066, https://doi.org/10.1016/j.rse.2024.114066, 2024.

Vuichard, N. and Papale, D.: Filling the gaps in meteorological continuous data measured at fluxnet sites with era-interim reanalysis, Earth Syst. Sci. Data, 7, 157-171, https://doi.org/10.5194/essd-7-157-2015, 2015.

Wang, Y., Meng, L., Liu, H., Luo, C., Bao, Y., Qi, B., and Zhang, X.: Construction and assessment of a drought-monitoring index based on multi-source data using a bias-corrected random forest (bcrf) model, Remote Sens., 15, 2477, https://doi.org/10.3390/rs15092477, 2023.

Yang, Y., Roderick, M. L., Guo, H., Miralles, D. G., Zhang, L., Fatichi, S., Luo, X., Zhang, Y., McVicar, T. R., Tu, Z., Keenan, T. F., Fisher, J. B., Gan, R., Zhang, X., Piao, S., Zhang, B., and Yang, D.: Evapotranspiration on a greening earth, Nat. Rev. Earth Environ., 4, 626-641, https://10.1038/s43017-023-00464-3, 2023.

Yao, Z., Li, W., and Cui, Y.: A_optram-et: An automatic optical trapezoid model for evapotranspiration estimation and its global-scale assessments, ISPRS J. Photogramm., 218, 181-197, 10.1016/j.isprsjprs.2024.10.019, 2024.

Zhang, G. and Lu, Y.: Bias-corrected random forests in regression, J. Appl. Statist., 39, 151-160, https://doi.org/10.1080/02664763.2011.578621, 2012.

Zhang, K., Kimball, J. S., and Running, S. W.: A review of remote sensing based actual evapotranspiration estimation, Wiley Interdiscip. Rev. Water, 3, 834-853, https://doi.org/10.1002/wat2.1168, 2016.

Zhang, Q., Liu, X., Zhou, K., Zhou, Y., Gentine, P., Pan, M., and Katul, G. G.: Solar-induced chlorophyll fluorescence sheds light on global evapotranspiration, Remote Sens. Environ., 305, 114061, https://doi.org/10.1016/j.rse.2024.114061, 2024.

Zhang, Y., Chiew, F., Zhang, L., Leuning, R., and Cleugh, H.: Estimating catchment evaporation and runoff using modis leaf area index and the penman-monteith equation, Water Resour. Res., 44, https://doi.org/10.1029/2007WR006563, 2008.

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

Zheng, C., Jia, L., and Hu, G.: Global land surface evapotranspiration monitoring by etmonitor model driven by multi-source satellite earth observations, J. Hydrol., 613, 128444, https://doi.org/10.1016/j.jhydrol.2022.128444, 2022.

Zhu, S., Clement, R., McCalmont, J., Davies, C. A., and Hill, T.: Stable gap-filling for longer eddy covariance data gaps: A globally validated machine-learning approach for carbon dioxide, water, and energy fluxes, Agric. For. Meteorol., 314, 108777, https://doi.org/10.1016/j.agrformet.2021.108777, 2022.