A benchmark dataset for half-hourly evapotranspiration estimation in China from 2000 to 2024

Long Qian[1,2], Xingjiao Yu[1,2], Lifeng Wu[1], Yaokui Cui[3], Zhitao Zhang[2], Junying Chen[2], Sumeng Ye[2], Xuqian Bai[2], Xiaogang Liu[4], Sien Li[5], Rangjian Qiu[6]

[1]School of Soil and Water Conversation, Jiangxi University of Water Resources and Electric Power, Nanchang, 330099, China
[2]College of Water Resources and Architectural Engineering, Northwest A&F University, Yangling, 712100, China
[3]Institute of RS and GIS, School of Earth and Space Sciences, Peking University, Beijing 100871, China
[4]Faculty of Modern Agricultural Engineering, Kunming University of Science and Technology, Kunming 650500, China
[5]Center for Agricultural Water Research in China, China Agricultural University, 100083, Beijing, China
[6]State Key Laboratory of Water Resources Engineering and Management, Wuhan University, 430072, Wuhan, China

Correspondence to: Lifeng Wu (china.sw@163.com)

Abstract. Latent heat flux (LE) provides a direct representation of terrestrial evapotranspiration (ET) and plays a critical role in hydrological cycle studies, land surface model development, and the evaluation of remotely sensed evapotranspiration products. Although flux observations based on the eddy covariance technique are widely regarded as essential benchmark data for evapotranspiration estimation, existing ChinaFlux observations are generally limited by short observation periods and extensive data gaps, which substantially constrain their applicability in long-term change analyses and multi-scale studies. To address these limitations, we developed a gap-filling and temporal prolongation framework specifically designed for half-hourly LE and established a continuous ground-based benchmark dataset covering China for the period 2000–2024 based on observations from 50 ChinaFlux sites. The framework is built upon an automated machine learning approach (AutoML) and integrates ERA5-Land reanalysis data with MODIS vegetation indices, enabling accurate gap-filling within observation periods and reliable prolongation beyond observation intervals. Comprehensive evaluations demonstrate that the AutoML framework achieves high accuracy at the half-hourly scale across different gap-length scenarios, with an overall correlation coefficient (CC) of 0.862 and a root mean square error (RMSE) of 33.75 W m^{-2}, and it substantially outperforms conventional methods under long-gap conditions (i.e., 7 d and 30 d) introduced in the artificial gap experiments. The forward and backward prolongation results show high consistency (CC values of 0.902 and 0.896, respectively) and exhibit robust temporal stability under varying training data lengths. Multi-timescale validations further indicate that the prolonged LE data reasonably reproduce diurnal variations, seasonal cycles, and interannual variability from half-hourly to daily and monthly scales. Comparisons with ChinaFlux observations under strict quality control (half-hourly LE observations with QA/QC=0) reveal good consistency across different temporal scales, underlying surface types, and climate zones. Interpretability analysis based on Shapley Additive Explanations (SHAP) indicates that energy supply consistently dominates LE variability, while vegetation state and water availability modulate their relative importance under different environmental conditions.

删除[Qianlong]: of H2O framework

Overall, we present the first continuous half-hourly ground-based LE benchmark dataset covering China for the period 2000–2024. This dataset provides essential data support for the evaluation of remotely sensed ET products, land surface model validation, and studies of regional water–energy cycles and climate change, and it is freely available via the following repository: https://doi.org/10.5281/zenodo.18194590 (Qian et al., 2026).

1 Introduction

Terrestrial evapotranspiration (ET) is a fundamental component of the land surface hydrological cycle, serving as the primary pathway by which water is transferred from the terrestrial surface to the atmosphere (Tang et al., 2024; Zhang et al., 2025). As a core process linking the atmosphere, hydrosphere, and biosphere, ET plays a pivotal role in regulating exchanges of water, carbon, and energy within the Earth system (Allen et al., 1998; Fisher et al., 2017). From an energy perspective, ET consumes approximately 48–62 % of the net solar radiation received at the land surface, thereby exerting a direct influence on surface cooling effects and cloud–precipitation feedbacks (Trenberth et al., 2009). From a hydrological perspective, more than 60 % of global terrestrial precipitation is returned to the atmosphere through ET, effectively constraining the natural upper limit of renewable freshwater resources (Oki and Kanae, 2006). Consequently, the availability of temporally continuous and quality-controlled ET benchmark data is a prerequisite for advancing our understanding of land–atmosphere interactions, evaluating evapotranspiration products derived from remote sensing and land surface models, and supporting water balance assessments and ecohydrological studies at basin and regional scales. In particular, at the site scale, flux observations based on the eddy covariance technique provide direct measurements of latent heat flux and its variability, and their temporal continuity and internal consistency are critical for model calibration, product validation, and cross-regional comparative analyses (Rodell et al., 2015; Cheng et al., 2021).

However, despite the central role of evapotranspiration in hydrological and ecosystem studies, existing observational approaches remain insufficient to fully meet the demands of regional- and long-term analyses. With the expansion of spatial and temporal domains and finer resolutions, land surface models, remote sensing retrievals, and data assimilation systems have become the primary means for generating spatially and temporally continuous ET estimates (Zhang et al., 2022; Yu et al., 2022; Zou et al., 2017). Nevertheless, the accuracy assessment and uncertainty characterization of these approaches rely heavily on calibration and validation against high-quality ground-based observations. Traditional ground-based measurement methods, such as evaporation pans, lysimeters, and the Bowen ratio technique, can provide relatively accurate estimates of ET or surface energy fluxes at the point scale (Allen et al., 2011). However, due to limitations in instrument deployment, maintenance requirements, and long-term operational stability, these methods are unable to form observation networks with broad spatial coverage, long-term continuity, and internal consistency (Xu and Singh, 2005). As a result, ground-based flux observations with high temporal resolution and strong physical consistency are particularly critical for regional and global-scale studies.

Within the existing ground observation systems, the eddy covariance (EC) technique has been widely regarded as the most important data source for constraining and evaluating ET models, as it directly measures turbulent exchanges between the land surface and the atmosphere. EC observations provide half-hourly measurements of latent heat flux (LE), while the associated meteorological or environmental variables are measured concurrently at the flux tower sites, thereby serving as an essential ground-based benchmark for evapotranspiration research. However, EC flux time series are commonly affected by substantial data gaps and long periods of missing observations due to instrument maintenance, complex meteorological conditions, and data transmission interruptions, with large variations in effective observation lengths among sites. In addition, EC measurements are subject to systematic uncertainties associated with energy balance non-closure, which may affect the accuracy of flux estimates even when observations are available. Taking the FLUXNET2015 dataset as an example, the average missing rate of hourly LE data is approximately 40 %, exceeding 70 % at some sites, and long gaps lasting more than 30 d account for a considerable fraction of the missing records (Foltýnová et al., 2020; Zhu et al., 2022). Although gap-filling is necessary to generate continuous LE time series, short gaps are often handled using various statistical or empirical approaches. Nevertheless, accurately reconstructing long and continuous LE time series remains challenging because evapotranspiration is governed by complex interactions among atmospheric conditions, vegetation dynamics, and water availability. Moreover, the available observation periods vary substantially among flux sites, and many sites do not provide sufficiently long records to support multi-decadal analyses. In addition, conventional gap-filling methods are primarily designed to reconstruct missing data within observation periods and are not intended for extending flux time series beyond the temporal coverage of available measurements, which further limits the applicability of flux datasets in long-term studies.

In China, the aforementioned limitations are also pervasive in ChinaFlux flux observations and are further exacerbated by insufficient site density, limited temporal continuity, and weak long-term consistency. Complex terrain, diverse climate regimes, and strong anthropogenic disturbances make flux tower measurements more susceptible to non-ideal conditions, resulting in frequent data gaps in latent heat flux (LE) time series and pronounced variability in data quality among sites. Although regional flux networks such as ChinaFlux have been established, the number of available flux tower sites in China has long been inadequate for regional- and global-scale ET product development and evaluation. Previous studies have shown that ET product validation over China often relies on fewer than 10 flux tower sites (Guo et al., 2022; Shi et al., 2024; Zuo et al., 2025), which is insufficient to represent the country's highly heterogeneous climate conditions and land surface characteristics. Even in comprehensive assessments based on global flux networks, Chinese sites remain underrepresented; for example, validation frameworks including nearly 200 global sites typically contain only 8–11 sites within China (Qian et al., 2023; Han et al., 2025; Zuo et al., 2025). In some data-driven modeling efforts and global ET product training datasets, the proportion of Chinese flux tower samples accounts for less than 2 % of the global total (Elnashar et al., 2021; Lu et al., 2021), which may be associated with multiple factors, including data availability and accessibility (Papale 2020), as well as the limited number and temporal continuity of flux sites, substantially limiting model applicability and generalization capability over China.

Although FLUXNET2015 has provided critical support for global evapotranspiration research (Jung et al., 2019; Qian et al., 2024a), its spatial coverage and temporal continuity within China remain limited. Recent studies indicate that global ET benchmark datasets derived from FLUXNET2015 include only a small number of Chinese sites and lack long-term continuous half-hourly flux time series for this region (Li et al., 2025a). Meanwhile, several systematic evaluation studies focusing on China have primarily been conducted at the monthly scale (Liu et al., 2025; Zuo et al., 2025), which hampers the characterization of ET dynamics at diurnal and short temporal scales. Under conditions of limited site availability and discontinuous observation records, discrepancies among different ET products over China are further amplified, with annual-scale estimates differing by up to nearly 50 % and exhibiting more pronounced regional contrasts in temporal trends (Cai et al., 2024). Against this background, there is an urgent need to systematically perform data gap-filling and temporal prolongation based on existing ChinaFlux observations and to establish a Chinese regional ground-based evapotranspiration benchmark dataset with extended temporal coverage and unified quality control standards. Such a dataset is essential for supporting the evaluation of remotely sensed ET products, the development of land surface and data-driven models, and studies of regional hydrological and ecohydrological processes.

Although model-based and remote sensing approaches are widely used for regional-scale evapotranspiration estimation, gap-filling and time-series reconstruction of flux tower latent heat flux (LE) observations still primarily rely on statistical and empirical methods. The marginal distribution sampling (MDS) method was first proposed and applied to flux data gap-filling by Reichstein et al. (2005), who systematically described the approach. Since then, it has become one of the standard gap-filling methods in eddy covariance flux data processing. Among these, MDS has been widely adopted as a standard approach due to its simplicity and reasonable performance for short gaps (Pastorello et al., 2020; Mahabbati et al., 2021; Vekuri et al., 2023). MDS fills missing flux data by identifying periods with similar meteorological conditions and using the mean or median of observed values, combining features of the lookup table (LUT) method and the mean diurnal variation (MDV) method. It has been widely applied in international flux networks such as FLUXNET and CarboEurope (Papale et al., 2006; Jia et al., 2015). However, because evapotranspiration is jointly controlled by energy availability, water supply, and vegetation dynamics, its responses are highly nonlinear and context dependent, which limits the ability of MDS—based on meteorological similarity assumptions—to represent complex causal relationships (Chen et al., 2012). Previous studies have shown that the performance of MDS decreases as gap length increases, primarily because its meteorological similarity assumptions cannot fully capture the nonlinear interactions among environmental drivers controlling evapotranspiration. As a result, reconstruction uncertainty tends to increase for prolonged missing periods, particularly when flux variability is strongly influenced by changing vegetation and hydrological conditions (Moffat et al., 2007; Kang et al., 2019; Zhu et al., 2022). Other conventional approaches, including mean diurnal variation, lookup table, and regression-based methods, exhibit similar limitations and are prone to systematic biases when data gaps are extensive or environmental conditions change markedly (Qian et al., 2024b).

Against this background, machine learning-based approaches have emerged as a promising alternative for LE gap-filling and flux time-series reconstruction at flux tower sites. Unlike traditional statistical and empirical methods, machine learning approaches can learn complex nonlinear relationships between LE and its meteorological and surface drivers directly from observations. These methods offer greater flexibility in handling long data gaps and complex environmental conditions when reference variables are continuously available (Kim et al., 2020; Zhu et al., 2022; Guo et al., 2022). Nevertheless, most conventional machine learning models require manually specified model structures and hyperparameter tuning, while deep learning approaches are often sensitive to sample size and data distribution, which can introduce instability when the number of available flux sites is limited (Khan et al., 2021). Automated machine learning (AutoML) has emerged as an efficient framework for automatically selecting algorithms, optimizing hyperparameters, and generating ensemble models through systematic model exploration (Gijsbers et al., 2019; Hutter et al., 2019; LeDell and Poirier, 2020). In particular, the H2O AutoML platform integrates multiple machine learning algorithms and stacked ensemble strategies, enabling robust predictive performance while substantially reducing manual intervention in model development (LeDell and Poirier, 2020; LeDell et al., 2021; Madni et al., 2023). Previous studies have demonstrated the effectiveness of AutoML for environmental and geoscientific applications characterized by complex nonlinear relationships and heterogeneous datasets (Ferreira et al., 2021; Bhatnagar et al., 2022; Bodini, 2023; Xu et al., 2025). When combined with reanalysis data featuring high temporal and spatial continuity, AutoML provides a flexible framework for reconstructing continuous flux time series and estimating latent heat flux beyond observation periods using relationships learned from available measurements and environmental drivers, thereby laying a methodological foundation for constructing long-term, continuous, site-level evapotranspiration benchmark datasets.

To address the widespread issues of extensive data gaps, inconsistent observation periods, and long-term sample scarcity in flux tower latent heat flux (LE) observations over China, we developed a site-level gap-filling and temporal prolongation framework for LE and, based on this framework, established a continuous evapotranspiration benchmark dataset for Chinese flux tower sites covering the period 2000–2024. Following rigorous quality control of EC observations and predefined site selection criteria (e.g., minimum observation length and data availability; see Sect. 2.1 for details), this study integrates observations from 50 ChinaFlux flux tower sites whose spatial distribution spans the major climate zones and representative underlying surface types across China, providing substantially improved site density and regional representativeness compared with previous studies. Methodologically, the framework is built upon an AutoML and combines reanalysis meteorological data with remotely sensed vegetation information to model the nonlinear relationships between LE and its driving factors at the site scale. This approach enables both high-accuracy gap-filling within observation periods and reasonable temporal prolongation beyond measurement intervals, resulting in continuous half-hourly LE time series.

Through a unified data processing workflow and consistent modeling strategy, LE data from all sites are harmonized in terms of temporal resolution and quality control standards, effectively alleviating the constraints imposed by long continuous data gaps and heterogeneous observation periods in the original measurements. To systematically assess the reliability and applicability of the constructed dataset, we designed a multi-level evaluation framework that includes gap-filling performance assessments under different gap-length scenarios, consistency tests between forward and backward temporal prolongation, and analyses of long-term temporal stability. These evaluations were further conducted across different underlying surface types and climate zones. The resulting dataset provides a temporally continuous, quality-controlled, and physically consistent ground-based evapotranspiration benchmark for China, offering essential data support for the evaluation of remotely sensed ET products, validation of reanalysis datasets and land surface process models, and studies of regional hydrological and ecosystem processes, while also establishing a robust data foundation for analyzing water–energy–carbon cycle dynamics over China under a changing climate.

2 Data and methodology

2.1 ChinaFlux

This study is based on eddy covariance (EC) observations from the ChinaFlux network, the primary long-term observation system for land–atmosphere interactions in China. Since 2002, ChinaFlux has been progressively established and operated across the country using standardized instrument configurations, observation protocols, and data processing procedures. The network systematically measures carbon, water, and energy fluxes over terrestrial ecosystems, together with accompanying meteorological variables, and provides one of the most important ground-based observational foundations for regional- and national-scale evapotranspiration (ET) studies in China. Within the ChinaFlux network, all flux towers employ the standard EC technique to continuously measure latent heat flux (LE). Raw high-frequency measurements are processed through a unified workflow to generate half-hourly flux products (Qi et al., 2021; Xu et al., 2023; Xu et al., 2024). This processing typically includes three-dimensional coordinate rotation, frequency response correction, Webb–Pearman–Leuning (WPL) correction, outlier removal, friction velocity (u*) filtering, and evaluation of energy balance closure, following quality control (QC) standards compatible with both ChinaFlux (Yu et al., 2006; Zheng et al., 2019) and FLUXNET (Pastorello et al., 2020). Previous studies have demonstrated that long-term observations from ChinaFlux sites are generally of reliable quality, with energy balance closure levels falling within internationally accepted ranges (Li et al., 2005; Wilson et al., 2002), thereby meeting the basic requirements for use as regional ET benchmark data.

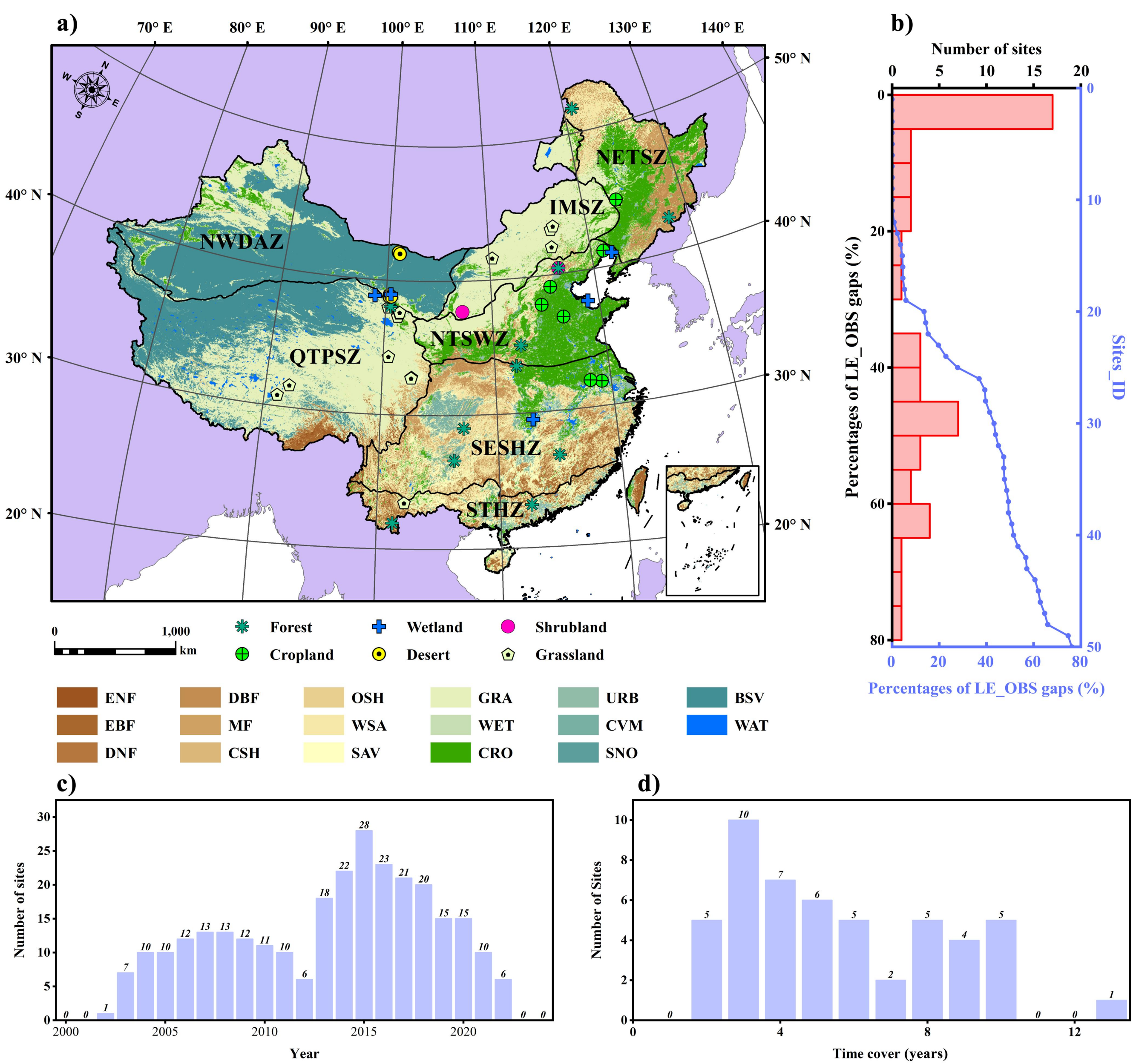
In total, this study integrates and selects 50 ChinaFlux flux tower sites (Fig. 1a), covering six major underlying surface types in China, including grassland (17 sites), shrubland (2 sites), cropland (9 sites), forest (12 sites), desert (5 sites), and wetland (5 sites). The vegetation cover categories shown in Fig. 1a were classified according to the IGBP land cover scheme, and the underlying land cover background was derived from the MODIS land cover product (MCD12Q1; Friedl et al., 2010). This land cover information was further cross-validated using the site-provided metadata to ensure consistency with the actual vegetation types at each flux tower. These sites also span seven major climate zones: the Northeast Temperate Subhumid Zone (NETSZ, 4 sites), Inner Mongolia Temperate Semiarid Zone (IMSZ, 5 sites), Northern Temperate Subhumid Warm Zone (NTSWZ, 9 sites), Southeast Subtropical Humid Zone (SESHZ, 7 sites), Southern Tropical Humid Zone (STHZ, 3 sites), Northwest Desert Arid Zone (NWDAZ, 10 sites), and Qinghai–Tibet Plateau Semiarid Zone (QTPSZ, 12 sites). The selected sites exhibit substantial heterogeneity in climatic conditions, vegetation types, and hydrothermal regimes, enabling a robust representation of the spatial variability of evapotranspiration processes across China. Detailed information on site locations, underlying surface types, climate zone classifications, and observation periods is provided in Table A1, and data source information for each flux tower can be seen in Table A2.

删除[Qianlong]: s

Given the limited number of flux tower sites and the scarcity of long-term continuous records in China, a relatively inclusive site selection strategy was adopted to maximize the use of available information while ensuring basic data reliability. Specifically, selected sites were required to meet the following criteria: (1) an effective observation period of at least 2 years, and (2) no fewer than 10 000 half-hourly LE records available for model training, ensuring a stable basis for gap-filling and temporal prolongation. Among the final set of sites, 40 provide quality-controlled original LE observations, while the remaining 10 sites only provide continuous LE time series that have already been gap-filled within their observation periods. For these 10 sites, the original observations and the investigator gap-filled data cannot be distinguished in the source records, and therefore the entire provided continuous series was retained as supplied by the data contributors. These gap-filled series are not considered "observations" in the dataset, and artificial gaps are introduced in the same manner as for the other 40 sites to ensure a consistent training strategy, and only temporally prolonged data beyond the observation period are generated. Within the observation period, no gap-filled data are created for these sites. For the other 40 sites, all investigator gap-filled values were strictly removed before our processing, and therefore no investigator gap-filled data are retained in the records used in this study. All sites are modeled independently, so the presence of these 10 sites does not affect the results or quality of the 40 original-observation sites. In the final dataset, each site corresponds to a separate file, and all data are clearly classified and labeled to distinguish different data sources and processing stages. Detailed descriptions of the data flags and file structure are provided in Section 5 ("Data availability"). These 10 sites are explicitly identified in Table A1 and in the dataset's Readme.txt file, allowing users to include or exclude them according to their specific research needs. Given the overall scarcity of ChinaFlux data, retaining these sites provides users with greater flexibility in selecting data for their specific research objectives.

删除[Qianlong]: . For these 10 sites,

The temporal coverage and data gap characteristics of the selected sites are also summarized. Fig. 1b illustrates the distribution of site-level missing data proportions of half-hourly LE, where the histogram (left axis) shows the number of sites within different gap-percentage intervals and the line (right axis) indicates the corresponding missing data percentage for each site (Site_ID as listed in Table A1). Fig. 1c presents the interannual variations in the number of available sites, while Fig. 1d shows the distribution of observation period lengths (in years) for all sites. All selected sites provide half-hourly LE data, which serve as the basis for subsequent gap-filling, temporal prolongation, and the construction of a continuous dataset spanning 2000–2024. By integrating these sites, this study assembles one of the most comprehensive flux tower–based evapotranspiration observation datasets for China to date in terms of site number and coverage of climate and underlying surface types, providing essential support for methodological evaluation, long-term consistency analyses, and the development of a regional ET benchmark dataset for China.



删除[Qianlong]:

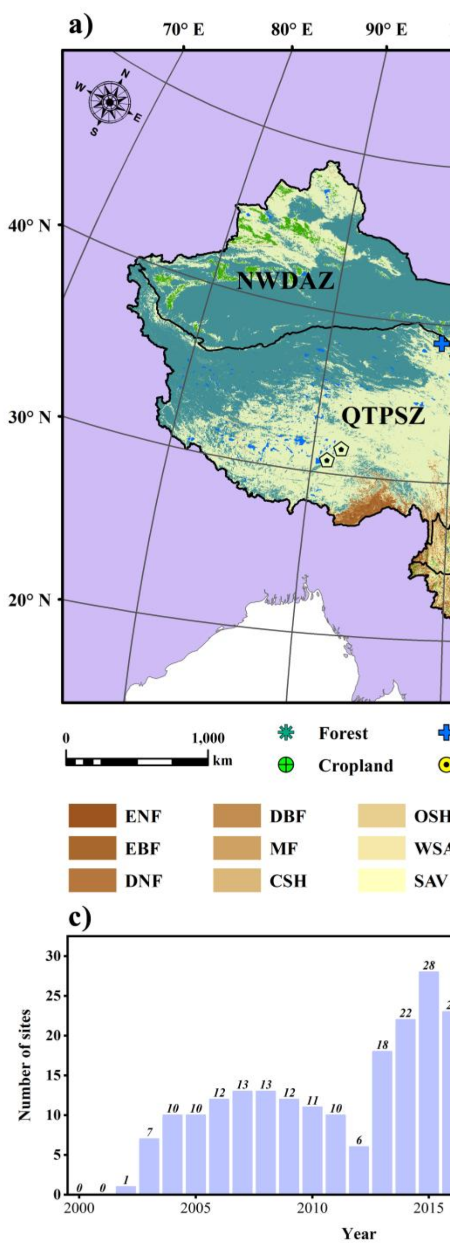


Figure 1: Spatial distribution of the selected ChinaFlux sites (a) and data coverage characteristics of half-hourly LE observations (b–d). In panel (b), the histogram (left y-axis) represents the number of sites within different percentages of LE observation gaps (top x-axis), while the blue line shows the distribution of gap percentages (bottom x-axis) for individual sites ordered by Site_ID (right y-axis). Panel (c) presents the number of available sites by year, and panel (d) summarizes the length of observation periods for all sites (years).

2.2 ERA5-Land data

To support gap-filling and temporal prolongation of flux tower latent heat flux (LE) observations, ERA5-Land reanalysis data were selected as site-scale meteorological and hydrological driving variables (Hersbach et al., 2020; Muñoz-Sabater et al., 2021). ERA5-Land is produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) and provides globally continuous coverage with high spatiotemporal consistency, offering a uniform background for flux reconstruction across different climate zones and land surface conditions (Muñoz-Sabater et al., 2021; Copernicus Climate Change Service, 2022). Hourly ERA5-Land data were extracted at each site location from the ECMWF/ERA5_LAND/HOURLY dataset using Google Earth Engine (GEE; https://code.earthengine.google.com/, last access: 11 November 2025). Bilinear interpolation was applied to derive representative values at the site scale, and all timestamps were converted from Coordinated Universal Time (UTC) to local time to ensure consistency with flux tower observations. To match the half-hourly temporal resolution of the EC measurements, hourly ERA5-Land variables were resampled to a half-hourly scale using linear interpolation. The ERA5-Land variables used in this study include latent heat flux (LE), surface air pressure (PA), precipitation (Pre), relative humidity (RH), surface net solar radiation (denoted here as Rn, following the ERA5-Land variable definition), downwards solar radiation (Rs), surface runoff (Runoff), volumetric soil water in the 0–7 cm layer (SMC_1) and volumetric soil water in the 7–28 cm layer (SMC_2), air temperature (Temp), vapor pressure deficit (VPD), and 10 m wind speed (WS). These variables collectively characterize key controls on evapotranspiration, including surface energy balance, atmospheric moisture and aerodynamic conditions, and soil moisture constraints.

2.3 MODIS data

To complement the meteorological drivers and better represent vegetation conditions, remotely sensed products from the Moderate Resolution Imaging Spectroradiometer (MODIS) were used to derive the normalized difference vegetation index (NDVI) (Didan, 2015) and leaf area index (LAI) (Myneni et al., 2015). NDVI data were obtained from the MODIS MOD13Q1 (Terra) and MYD13Q1 (Aqua) products, while LAI data were derived from the MODIS MOD15A2H and MYD15A2H products. All products were converted to physical values using the corresponding scale factors provided in the official product user guides, with NDVI scaled by 0.0001 (Didan and Barreto Munoz, 2019) and LAI scaled by 0.1° (MODIS LAI/FPAR Product Team, 2020). NDVI and LAI data were extracted at the site scale using GEE (https://code.earthengine.google.com/, last access: 11 November 2025), with bilinear interpolation applied for spatial sampling. The spatial resolutions of the NDVI and LAI products are 250 m and 500 m, respectively. Given the relatively low temporal resolution of MODIS products (i.e., 16 day composites for NDVI and 8-day composites for LAI), NDVI and LAI were assumed to remain constant within each compositing period, and the corresponding values were assigned uniformly to all half-hourly (and daily) time steps within that period to ensure temporal alignment with the half-hourly LE time series. This assumption is commonly adopted in flux–remote sensing integration studies and is generally considered reasonable because MODIS compositing procedures are designed to reduce short-term noise (e.g., cloud contamination and atmospheric effects) and to represent average vegetation conditions over the compositing interval (Didan et al., 2015; Myneni et al., 2002).

It should be noted that MODIS NDVI and LAI products are available from 18 February 2000 at 08:00 (local time). For periods prior to this date, NDVI and LAI data are unavailable. In these cases, vegetation indices were treated as missing values and were not included as predictors in model training and prediction. Given that the AutoML framework is primarily based on tree-based algorithms capable of handling missing values, LE estimation during this early period relied on meteorological and energy-related drivers.

2.4 Generation of gap-filling

To construct temporally continuous latent heat flux (LE) time series suitable for long-term consistency analyses, a data-driven gap-filling framework was implemented to reconstruct missing LE observations at the flux tower scale. Given the substantial heterogeneity among flux sites in terms of climate background, underlying surface type, and observation length, a site-specific modeling strategy was adopted. Each flux tower site was treated as an independent modeling unit, for which models were separately trained, evaluated, and applied to generate seamless half-hourly LE time series over the study period. During model development and evaluation, the original LE observations were first subjected to quality screening, following the standard quality control procedures of FLUXNET (Pastorello et al., 2020) and ChinaFlux (Yu et al., 2006; Zheng et al., 2019). Specifically, only observations flagged as high quality (QC/QA = 0) were retained, while data affected by instrument malfunction, low turbulence conditions (e.g., insufficient u*), or energy balance non-closure were excluded. Only reliable records were retained for model training. To systematically assess model performance under different gap conditions, multiple artificial gap scenarios were constructed along the time dimension by deliberately removing data segments of varying lengths from the complete observation records following the widely adopted artificial-gap evaluation strategy used in flux gap-filling studies (Moffat et al., 2007; Li et al., 2025). This design enables a robust evaluation of gap-filling performance across different temporal scales. In addition, conventional approaches widely used in flux studies were included as benchmark methods, allowing quantitative comparisons with data-driven approaches in terms of accuracy and stability (Moffat et al., 2007). The workflow for generating the final half-hourly LE data products is illustrated in Fig. 2a, while the workflows for testing the gap-filling and prolongation algorithms are shown in Fig. 2b and Fig. 2c, respectively.

删除[Qianlong]: The overall technical workflow and evaluation framework are illustrated in Fig. 2.

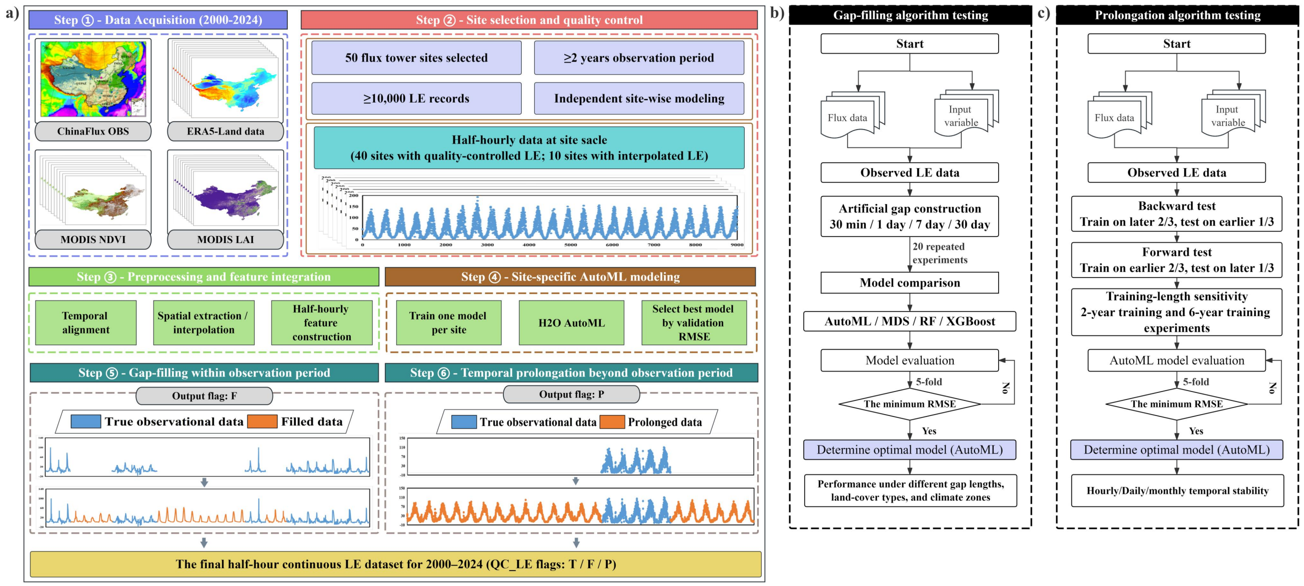
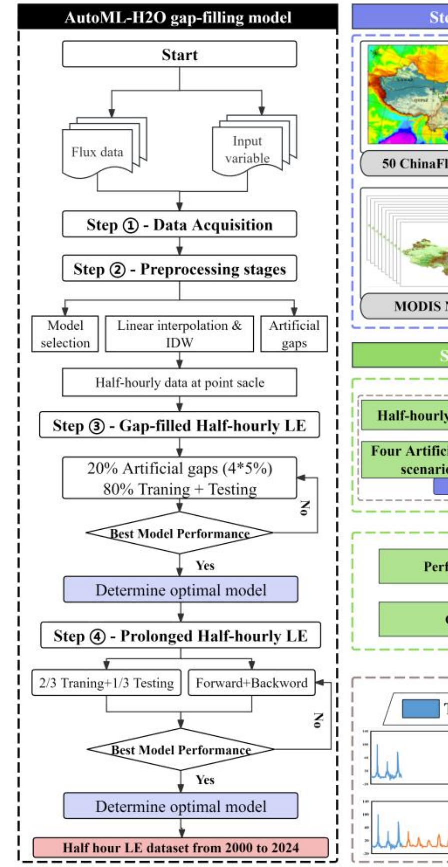


Figure 2: Workflow for generating the final half-hourly LE data products from 2000 to 2024 (a). Workflows used to test the gap-filling (b) and prolongation algorithms (c). The evaluation workflows were independent from the production workflow. Artificial gaps and forward/backward temporal splits were used only to assess model performance, whereas the final released products were generated by retraining the site-specific AutoML models using all available high-quality LE observations at each site.

2.4.1 AutoML

AutoML of H2O framework (LeDell and Poirier, 2020; H2O.ai, 2023) was adopted as the core modeling tool for LE gap-filling in this study. H2O AutoML has been widely used for regression and prediction tasks in environmental and geoscientific applications (Guo et al., 2024; Li et al., 2025b; Zhao et al., 2026). Unlike conventional machine learning approaches that require manual specification of model types and hyperparameters—often leading to high tuning costs and reduced transferability across sites—AutoML automatically performs model selection, hyperparameter optimization, and model ranking within a predefined search space (LeDell and Poirier, 2020). This strategy improves modeling efficiency and robustness across multiple sites while maintaining reproducibility. For each flux tower site, an independent AutoML regression model was constructed, with site-level LE observations as the target variable and a set of multi-source environmental drivers derived from ERA5-Land reanalysis data and MODIS vegetation indices as input features. These predictors were used to capture the nonlinear relationships between LE and its meteorological, hydrological, and vegetation controls.

During the AutoML search process, multiple commonly used regression algorithms, including tree-based models and their ensemble variants, were evaluated, and the optimal model was automatically selected based on validation performance. Here, AutoML was not treated as a single predefined algorithm, but as an automated modeling framework that searches across multiple candidate algorithms and hyperparameter configurations. In parallel, the benchmark machine learning models were also tuned using the same validation-based principle: different parameter combinations were iteratively tested, and the



删除[Qianlong]:

删除[Qianlong]: **Flowchart and schematic of the gap-filling and prolongation framework for LE_OBS data.**

删除[Qianlong]: , which has been widely used parameter setting with the lowest validation root mean square error (RMSE) was retained as the optimized model for that method. After the optimized AutoML and benchmark models were determined, their performance was compared, and the best-performing modeling approach was selected for generating the final data product. Because AutoML consistently showed superior or comparable performance in almost all evaluation scenarios, it was ultimately adopted for producing the gap-filled and prolonged LE time series.

Specifically, the H2O AutoML framework evaluates a suite of candidate algorithms, including gradient boosting machine (GBM), distributed random forest (DRF), extreme gradient boosting (XGBoost), generalized linear models (GLM), and deep learning neural networks (DNN), as well as stacked ensemble models that combine multiple base learners. During the automated search process, multiple model configurations and hyperparameter combinations are explored through a random grid search strategy, and models are ranked based on cross-validation and validation performance. In this study, RMSE on the validation set was used as the criterion for defining model optimality. For each flux tower site, two site-specific final models were selected from the H2O AutoML leaderboard: one model for gap-filling within the observation period and another model for temporal prolongation beyond the observation period. Each final model corresponded to the candidate model with the lowest validation RMSE among all evaluated algorithms and hyperparameter configurations for its respective task. After selection, these two models were fixed for the corresponding site and task, and the selected algorithms did not change over time within the record of a given flux tower. In this study, no strict limit was imposed on the number of candidate models (i.e., the number of models is adaptively determined by the AutoML process), allowing the framework to fully explore the model space and select the optimal model for each flux tower site.

2.4.2 Artificial gap scenarios

In flux tower latent heat flux (LE) observations, data gaps vary substantially in duration, ranging from single missing time steps to continuous gaps lasting several weeks or longer. To systematically evaluate model performance across different missing-data scales, four artificial gap scenarios were constructed within the observation period of each site, corresponding to continuous gap lengths of 30 min, 1 d, 7 d, and 30 d. These four scenarios were used only for performance evaluation and method comparison, not as four independent production gap-filling procedures. For each scenario, the artificially removed data accounted for approximately 5 % of the total valid observations at the site, ensuring comparability of evaluation results across gap-length conditions. Artificial gaps were generated using a sliding-window approach to create continuous missing segments. Only time windows in which the proportion of valid original observations exceeded 50 % were considered eligible, thereby avoiding the introduction of artificial gaps within periods of inherently poor data quality. Artificial gaps generated under different scenarios did not overlap in time. For the 30 min scenario, single half-hourly observations were randomly removed, with the additional constraint that valid observations were present immediately before and after the removed record to preserve the basic continuity of the time series. During model training and validation, the remaining data after artificial gap removal were randomly divided into a training set (80 %) and a validation set (20 %). Five-fold cross-validation was applied during training to evaluate model performance. For each data split, the model parameters yielding the

删除[Qianlong]:

删除[Qianlong]: root mean square error (

删除[Qianlong]:)

删除[Qianlong]: The final model for each site was selected as the leader model from the H2O AutoML leaderboard, corresponding to the candidate model with the lowest validation RMSE among all evaluated algorithms and hyperparameter configurations. The final model for each site was selected as the leader model from the AutoML leaderboard, which represents the best-performing model (including possible ensemble models) under the given validation metrics.

best validation performance were retained and subsequently applied to fill the corresponding artificial gaps, allowing assessment of gap-filling accuracy and stability under different missing-data scales. This procedure was conducted independently for each site and repeated 20 times using different random data splits and gap realizations to reduce the influence of randomness. For each site, artificial gap scenario, and gap-filling method, the performance metrics were first calculated separately for each of the 20 repeated experiments and then averaged across the 20 repetitions to obtain one site-level value for each metric. Therefore, in the cross-site comparison shown in Fig. 3, each boxplot is constructed from 50 site-level averaged values, with each flux tower contributing one value. After this evaluation step, the artificial gaps were not used in the final data product. Instead, a single final AutoML model was retrained for each site using all available high-quality LE observations and then applied once to fill the actual missing LE values within the observation period. This process resulted in seamless half-hourly LE time series, which serve as the baseline dataset for subsequent temporal prolongation and multi-scale analyses.

To facilitate comparative evaluation of different gap-filling approaches, four independent methods were considered in this study: AutoML, random forest (RF), extreme gradient boosting (XGBoost), and marginal distribution sampling (MDS). AutoML was treated as an automated modeling framework that searches over multiple internal candidate algorithms and hyperparameter configurations and then selects the best-performing leader model based on validation RMSE; it does not select from among the benchmark methods used for comparison in this study. The traditional marginal distribution sampling (MDS) method (Foltýnová et al., 2020) and two commonly used machine learning algorithms, RF (Khan et al., 2021) and XGBoost (Liu et al., 2025), were implemented as benchmark methods. These three benchmark methods were trained and evaluated independently of AutoML, and therefore do not belong to the AutoML framework. These benchmark methods were also applied only within the artificial-gap evaluation framework and were not used to generate the final released gap-filled LE dataset. Model performance was assessed using five-fold cross-validation during the training stage to optimize model configuration and reduce the risk of overfitting. After model selection, the optimal model for each site was retrained using all available high-quality LE observations and then applied to fill the missing LE values. For the final released dataset, the site-specific AutoML model was used for production, rather than MDS, RF, or XGBoost. Consequently, a site-specific gap-filling model was established for each site, providing a reliable basis for subsequent temporal prolongation of LE time series.

It should be noted that most ChinaFlux sites do not provide complete sets of meteorological driving variables (e.g., shortwave radiation, air temperature, and vapor pressure deficit). Therefore, ERA5-Land variables were consistently used as reference inputs for all benchmark methods to ensure that different algorithms were compared under identical information conditions. Due to the lack of complete and continuous in situ meteorological observations at many flux tower sites, a comprehensive site-by-site comparison between ERA5-Land data and locally measured variables could not be conducted across all sites. However, for a subset of sites where both ERA5-Land data and in situ meteorological observations are available, we found that the agreement between ERA5-Land and site measurements is generally good, with coefficients of determination (R^2) typically ranging from approximately 0.70 to 0.90 for key variables such as air temperature, radiation,

删除[Qianlong]: t

删除[Qianlong]: random forest (

删除[Qianlong]:)

删除[Qianlong]: extreme gradient boosting (

删除[Qianlong]:)

and VPD. This practice has been widely adopted in flux data gap-filling studies and provides a reasonable representation of site-scale meteorological backgrounds when in situ measurements are unavailable. Although the use of reanalysis data may introduce additional uncertainty compared to in situ observations, employing ERA5-Land ensures spatial and temporal consistency of input variables across all sites, which is essential for developing a unified modeling framework and enabling fair comparisons among different methods. The MDS method performs gap-filling based on meteorological similarity assumptions, whereas RF and XGBoost predict LE by learning statistical relationships between LE and multi-source environmental drivers. All benchmark methods were implemented under the same artificial gap scenarios and data partitioning schemes as AutoML, ensuring fairness and reproducibility in performance comparisons.

In addition, three commonly used performance metrics were employed to quantify gap-filling performance, including the correlation coefficient (CC), root mean square error (RMSE, W m^{-2}), and percentage bias (PBias, %). The definitions of CC and RMSE follow Li et al. (2025a), while PBias was calculated according to Qian et al. (2023). RMSE was used as the model selection criterion, whereas CC and PBias were reported for supplementary performance evaluation. In addition to these quantitative metrics, we further evaluated the temporal behavior of the reconstructed LE time series at representative sites. This assessment examined whether the reconstructed LE time series preserved the observed diurnal patterns, amplitude variations, and temporal continuity during extended missing periods. The purpose of this analysis was not to provide an additional model-selection metric, but rather to evaluate the physical plausibility and temporal consistency of the reconstructed time series, which may not be fully reflected by pointwise statistical metrics alone. Similar visual assessments of temporal structure have been adopted in recent flux gap-filling studies (Li et al., 2025). Therefore, the diel-cycle evaluation was used only as a complementary assessment of reconstruction quality. Finally, two descriptive statistics were calculated, including the mean LE and coefficient of variation (CV, %), these statistics were used to quantitatively compare the ability of different gap-filling methods to reproduce the central tendency and temporal variability of the observed LE series during long continuous gaps.

删除[Qianlong]: And

删除[Qianlong]: Except to these quantitative metrics, a qualitative assessment of diel-cycle reconstruction was conducted for representative sites.

2.4.3 Prolonged hourly LE data

After completing seamless gap-filling within the observation period, the latent heat flux (LE) time series at each flux tower site were further temporally prolonged at the half-hourly scale to construct continuous records covering a unified time span. Temporal prolongation was also implemented within the AutoML of H2O framework, using the same model architecture, input feature set, and training strategy as described in Sect. 2.4.1 to ensure methodological consistency. For each site, a site-specific prolongation model was independently trained using the available LE observations as supervisory information. Five-fold cross-validation was applied to identify the optimal model configuration and hyperparameter combination. Temporal prolongation was performed in two directions at the half-hourly scale. For a site with an observation period spanning a specific interval (e.g., 2006–2015), the period preceding the observations (e.g., 2000–2005) was defined as backward prolongation, whereas the period following the observations (e.g., 2016–2022) was defined as forward prolongation. Under identical data and modeling conditions, the prolongation performance in both temporal directions is expected to be comparable. To evaluate the consistency of prolongation results across temporal directions, a symmetric data partitioning strategy was adopted. For backward prolongation, the last two-thirds of the observed time series were used for model training, while the first one-third served as the testing set. Conversely, for forward prolongation, the first two-thirds of the data were used for training and the remaining one-third for testing. Model performance on the testing sets was evaluated using CC, RMSE, and PBias. Optimal prolongation models were selected based on the lowest RMSE, while CC and PBias served as complementary metrics to assess correlation and systematic bias. This approach enables an objective evaluation of directional consistency and the detection of potential systematic deviations during temporal prolongation.

In addition, to examine the temporal stability of model performance as the length of the prolongation period increases, two representative training-length scenarios were designed. For sites with observation periods of at least 6 years, the first 6 years of data were used for training and the remaining years for testing. For sites with observation periods of at least 2 years, the first 2 years were used for training, with subsequent years reserved for testing. By comparing prolongation performance under different training-length conditions, using the same metrics (CC, RMSE, PBias) and selecting models based on RMSE, the long-term stability and reliability of the model during temporal extension were systematically evaluated. All prolongation and validation procedures were conducted independently at the site scale. After completing consistency and stability assessments, the final model for each site was retrained using all available high-quality LE observations and applied to periods outside the observation interval, generating continuous half-hourly LE time series covering 2000–2024.

2.5 Generation of LE data

Based on the gap-filling and temporal prolongation procedures described above, seamless site-level half-hourly latent heat flux (LE) time series were finally constructed covering the period from 00:00 on 1 January 2000 to 23:30 on 31 December 2024. The resulting dataset is temporally continuous, processed using a unified workflow across all sites, and explicitly distinguishes data generated through different procedures. At the half-hourly scale, LE records are assigned one of three quality flags according to their data source. The flag T denotes half-hourly LE values directly derived from original flux tower observations. The flag F represents half-hourly LE values obtained by gap-filling missing observations within the measurement period using the AutoML framework. The flag P indicates half-hourly LE values generated through temporal prolongation beyond the observation period based on site-level models. These quality flags enable users to flexibly select datasets consisting solely of observed data or to include gap-filled and prolonged records depending on specific research objectives.

On this basis, the half-hourly LE data were further aggregated to generate daily, monthly, and annual LE datasets spanning 2000–2024. Daily LE values were obtained by aggregating half-hourly records, while monthly and annual values were subsequently derived from the corresponding daily datasets. At aggregated temporal scales, a single categorical quality flag (T/F/P) was not assigned to each value, as aggregated data typically consist of a mixture of observed and reconstructed (gap-filled or prolonged) records. Instead, a quantitative indicator, N_obs, was provided to represent the number of half-hourly observations (QC_LE = T) contributing to each aggregated value (e.g., N_obs ranges from 0 to 48 for daily data). LE products at all temporal scales share the same time coverage and are linked to the half-hourly quality information through N_obs, ensuring consistency and traceability across multi-temporal-scale products.

3 Result

3.1 Evaluation of half-hourly gap-filled LE data

3.1.1 Gap-filling performance under four artificial gap scenarios

Under the four artificial gap-length scenarios (30 min, 1 day, 7 day, and 30 day), the evaluated methods exhibit pronounced differences in half-hourly LE gap-filling performance (Fig. 3). Overall, AutoML consistently achieves the best and most stable performance across all gap scales, with an average correlation coefficient (CC) of 0.862 and a root mean square error (RMSE) of 33.75 W m^{-2}. This performance substantially exceeds that of the conventional marginal distribution sampling (MDS) method (CC = 0.687) and is accompanied by a more concentrated bias distribution. For the shortest gap scenario (30 min), all methods are generally able to reconstruct short-term LE variability. Nevertheless, AutoML still yields the highest correlation (CC = 0.910) and the lowest error (RMSE = 30.52 W m^{-2}), reducing RMSE by 4.75 W m^{-2} relative to MDS. The performances of RF and XGBoost fall between AutoML and MDS, with overall biases close to zero. As gap length increases, performance differences among methods become progressively more pronounced.

Under the 1 day and 7 day scenarios, MDS performance deteriorates continuously, whereas AutoML maintains relatively high accuracy and stable bias control, indicating strong adaptability to moderate-length continuous gaps. The divergence among methods is most evident under the most extreme scenario of 30 d continuous gaps. In this case, AutoML still achieves a CC of 0.819 and an RMSE of 37.26 W m^{-2}, whereas the RMSE of MDS increases markedly to 49.85 W m^{-2}. Although RF and XGBoost outperform MDS under this scenario, their overall accuracy and stability remain inferior to those of AutoML.Overall, gap-filling performance exhibits a clear dependence on gap length. The MDS approach is only suitable for short gaps and shows rapid performance degradation as gap length increases. In contrast, AutoML maintains higher correlations, lower errors, and more stable bias characteristics across all gap-length scenarios, with particularly pronounced advantages under long continuous gaps. These results indicate that AutoML is better suited for constructing long-term, continuous half-hourly LE time series.

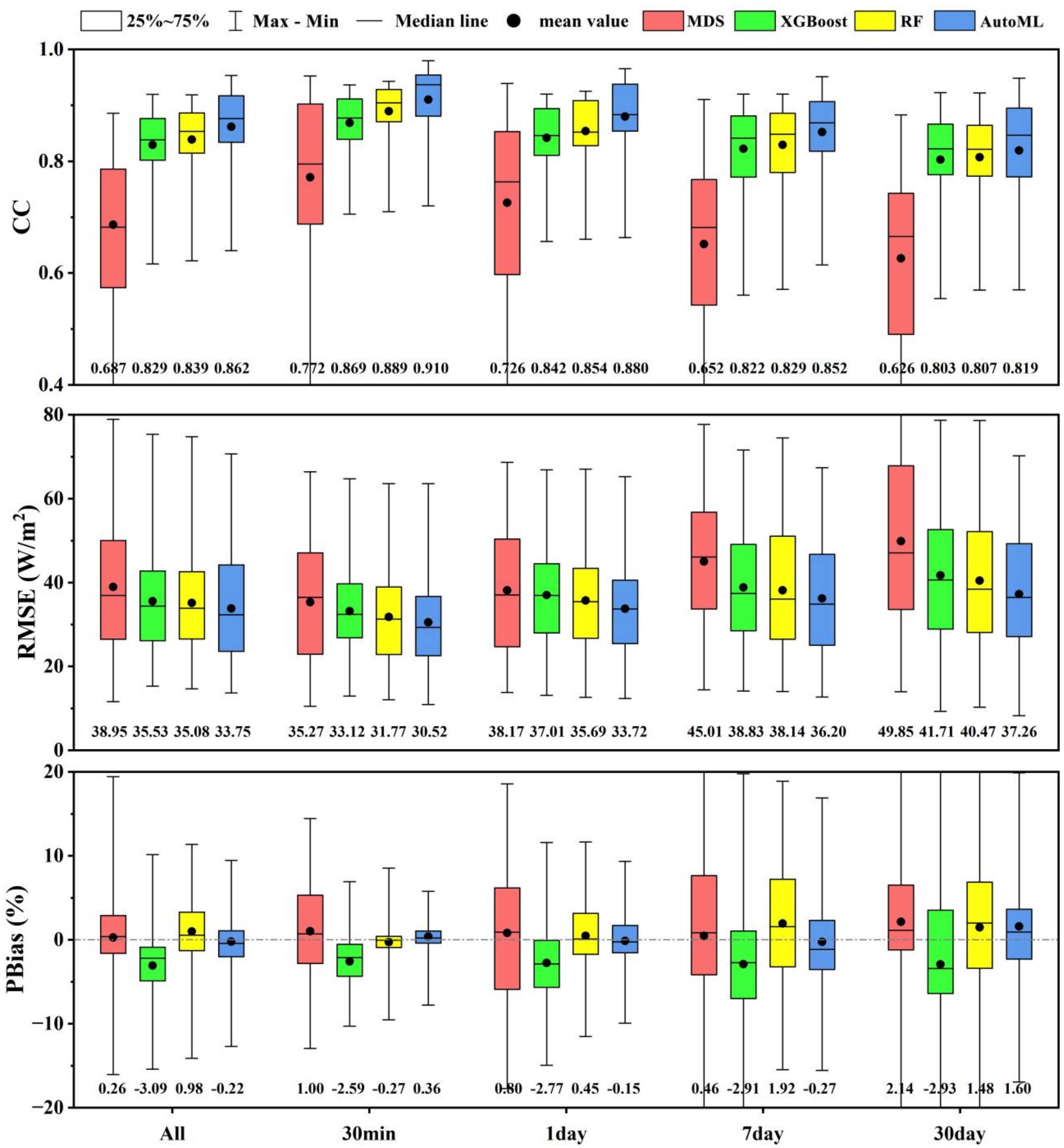


Figure 3: Comparison of half-hourly LE gap-filling performance among different methods under various artificial gap scenarios. **Each boxplot summarizes the distribution of site-level performance across the 50 flux tower sites for a given method and artificial gap scenario. For each site, artificial gaps were generated and filled in 20 repeated experiments, and the resulting metric values were first averaged across the 20 repetitions. Thus, each site contributes one averaged value to each boxplot, and each boxplot is constructed from 50 site-level averaged values. The evaluated metrics include correlation coefficient (CC), root mean square error (RMSE), and percent bias (PBias). The box indicates the interquartile range (25th–75th percentiles), the horizontal line denotes the median, the whiskers show the minimum and maximum values, and the point denotes the mean across the 50 site-level values. The numbers shown above or below the boxes indicate the corresponding mean values.**

删除[Qianlong]: **For each flux tower, each metric (CC, RMSE, PBias) represents the mean value across 20 repeated experiments. Each site thus contributes a single averaged value for each metric in the figure.**

Across different underlying surface types (Fig. 4a–c), all methods exhibit a common pattern of decreasing gap-filling performance with increasing gap length, with the most pronounced degradation occurring under the 30 day continuous gap scenario. In terms of overall differences among underlying surface types, cropland and wetland show the best gap-filling performance, maintaining relatively high correlations and low errors across all gap-length scenarios (e.g., AutoML achieves overall CC values of 0.885 and 0.894 for cropland and wetland, respectively). Forest and grassland exhibit intermediate performance, with gap-filling accuracy and stability falling between the high- and low-performing underlying surface types. In contrast, desert and shrubland represent the most challenging underlying surface types for gap-filling. Except for AutoML, the other methods generally show larger errors, pronounced systematic biases, and reduced stability under these conditions. AutoML maintains a relatively concentrated PBias distribution and lower RMSE even under these unfavorable conditions, indicating stronger robustness. Results across different climate zones (Fig. 4d–f) reveal a consistent dependence on gap length, with all methods experiencing the most substantial performance degradation under the 30 day scenario. In terms of regional differences, gap-filling performance is relatively better in the Inner Mongolia Temperate Semiarid Zone (IMSZ) and the Northeast Temperate Subhumid Zone (NETSZ), characterized by higher correlations and lower errors. For example, AutoML achieves an overall CC of 0.839, RMSE of 22.69 W m^{-2}, and PBias of 0.50 % in IMSZ, while NETSZ exhibits the highest overall CC (0.910) with a relatively low RMSE (28.28 W m^{-2}). In contrast, the Southeast Subtropical Humid Zone (SESHZ), Southern Tropical Humid Zone (STHZ), and Northern Temperate Subhumid Warm Zone (NTSWZ) generally show lower CC values and higher RMSE values, indicating reduced reconstruction accuracy under these climatic conditions. However, although arid and semiarid regions generally exhibit stronger correlations and lower RMSE values, some regions, such as the Northwest Desert Arid Zone (NWDAZ), show comparatively larger systematic deviations, with PBias reaching. Overall, the results indicate that model performance varies across climate zones, with different regions exhibiting distinct strengths and limitations when assessed from the perspectives of correlation, absolute error, and systematic bias.

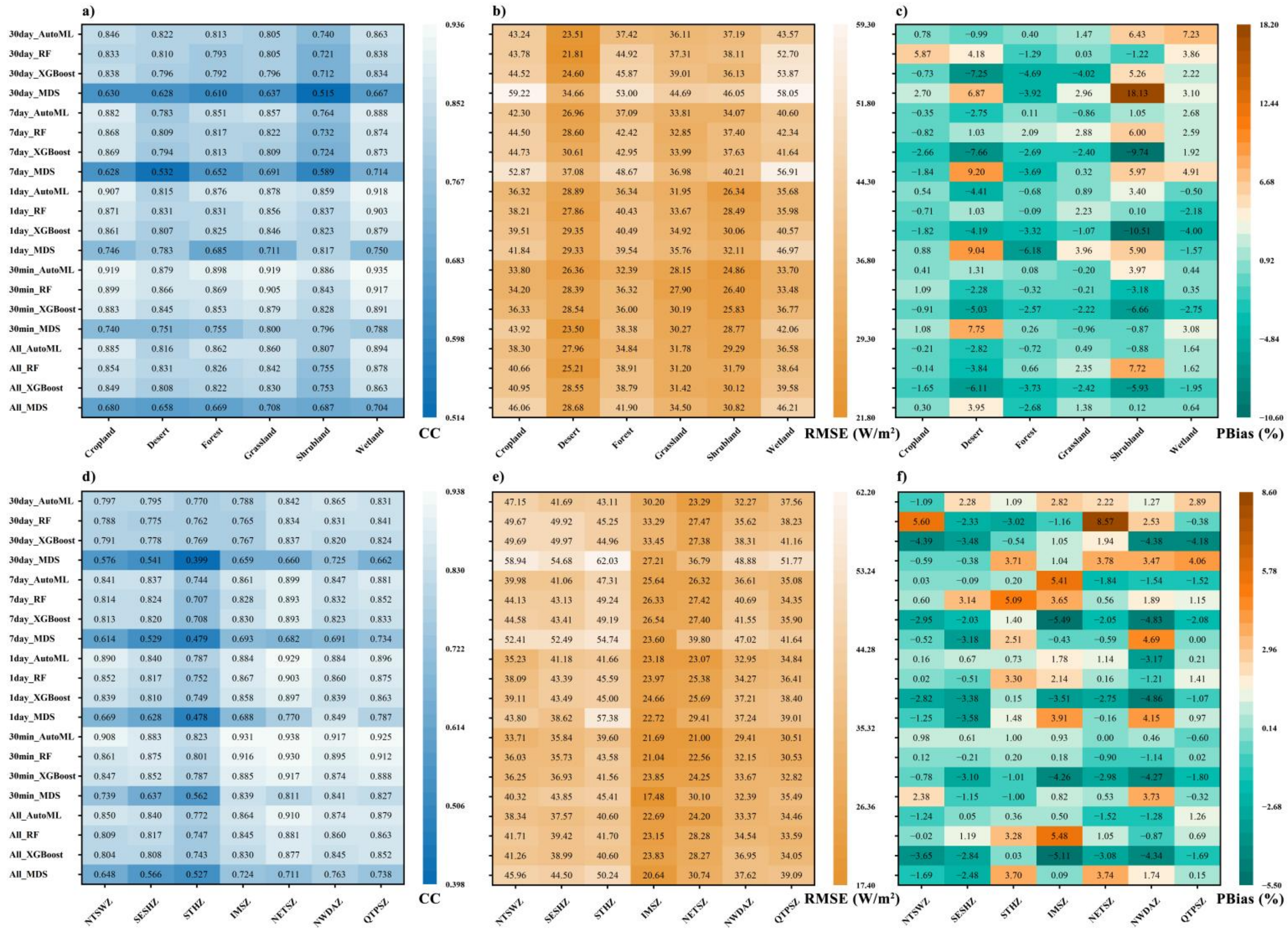


Figure 4: Performance of different methods across underlying surface types and climate zones under varying artificial gap scenarios. The results shown here represent evaluation outcomes under different artificial gap-length scenarios and do not correspond to separate released gap-filled datasets.

删除[Qianlong]: **The final released dataset was generated using one site-specific optimal model for each flux tower site.**

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

Under the artificially imposed 30 day continuous gap scenario, the evaluated gap-filling methods exhibit pronounced differences in their ability to reconstruct half-hourly LE time series (Figs. 5 and 6). Across different underlying surface types, AutoML consistently reproduces the diurnal cycles and amplitude characteristics of LE at cropland, forest, grassland, and wetland sites. Even during extended missing periods, the gap-filled results remain in good agreement with the observed time series. The daily mean LE and coefficient of variation (CV) further support this conclusion. Across all representative sites, AutoML generally produces mean LE values closer to observations than MDS and better preserves temporal variability. For example, at the forest site CN-ARF, the observed mean LE is 81.7 W m^{-2}, compared with 79.0 W m^{-2} for AutoML and 72.1 W m^{-2} for MDS. Similar improvements are observed across most underlying surface types.

In contrast, MDS frequently exhibits pronounced overestimation or underestimation across multiple underlying surface types, leading to weakened or distorted diurnal structures; this issue is particularly evident at desert and shrubland sites. Results across different climate zones show a similar pattern. In arid and high-altitude regions such as the NWDAZ and QTPSZ, AutoML maintains coherent and continuous temporal patterns during the gap period, whereas MDS displays larger fluctuations and reduced stability. In humid and subtropical climate zones, including NETSZ, NTSWZ, SESHZ, and STHZ, overall uncertainty increases; nevertheless, AutoML is still able to capture the primary temporal structure of LE, while MDS tends to amplify noise and attenuate diurnal signals. The statistical summaries shown in Fig. 6 also indicate that AutoML generally reproduces both the mean state and variability of LE more accurately than MDS across different climate zones. Overall, these representative examples demonstrate that under long continuous gap conditions, AutoML more effectively preserves the physical plausibility and temporal consistency of LE time series across different underlying surface types and climate zones. This capability substantially reduces the risk of abnormal fluctuations and structural distortions in the reconstructed LE records.

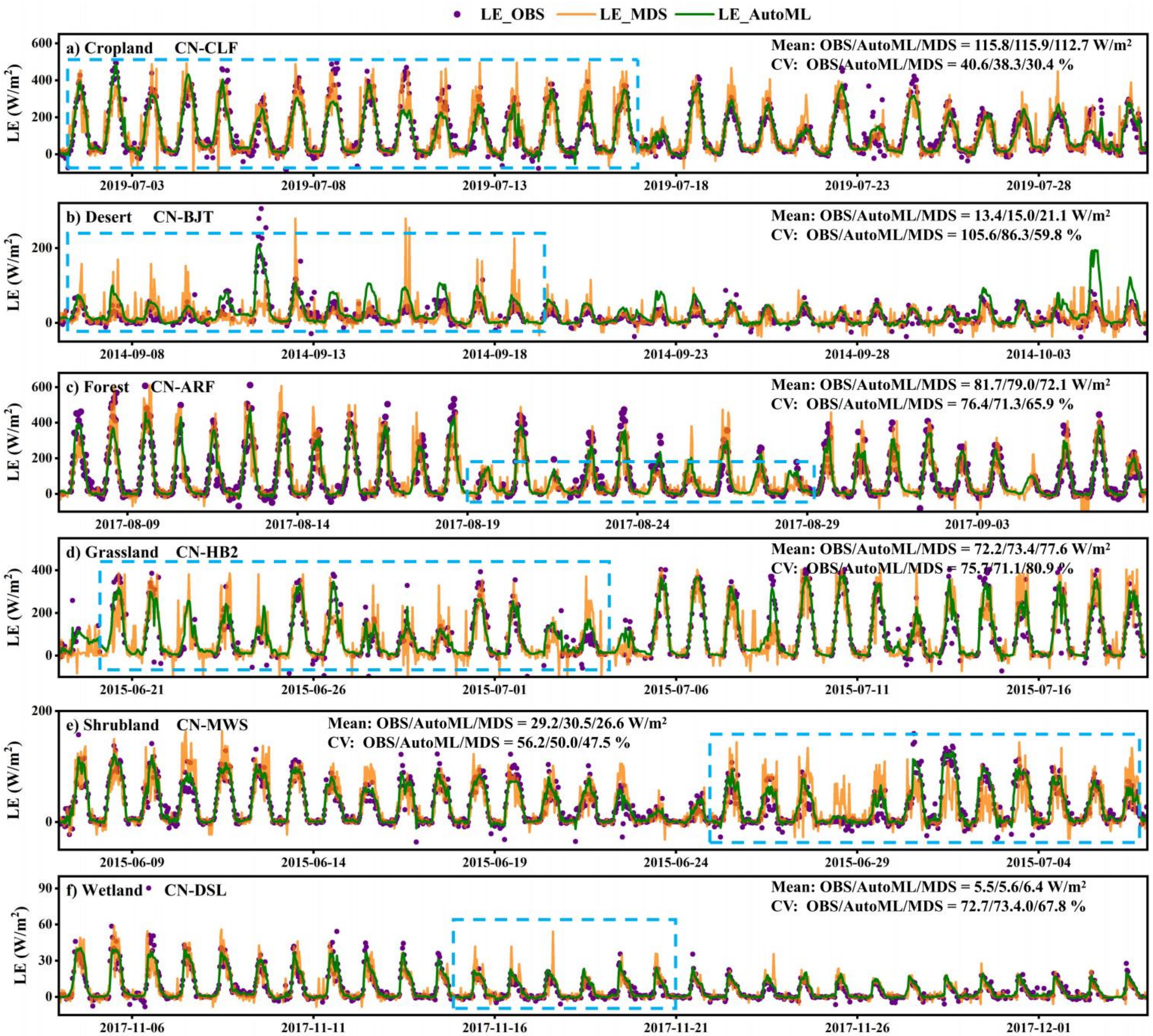
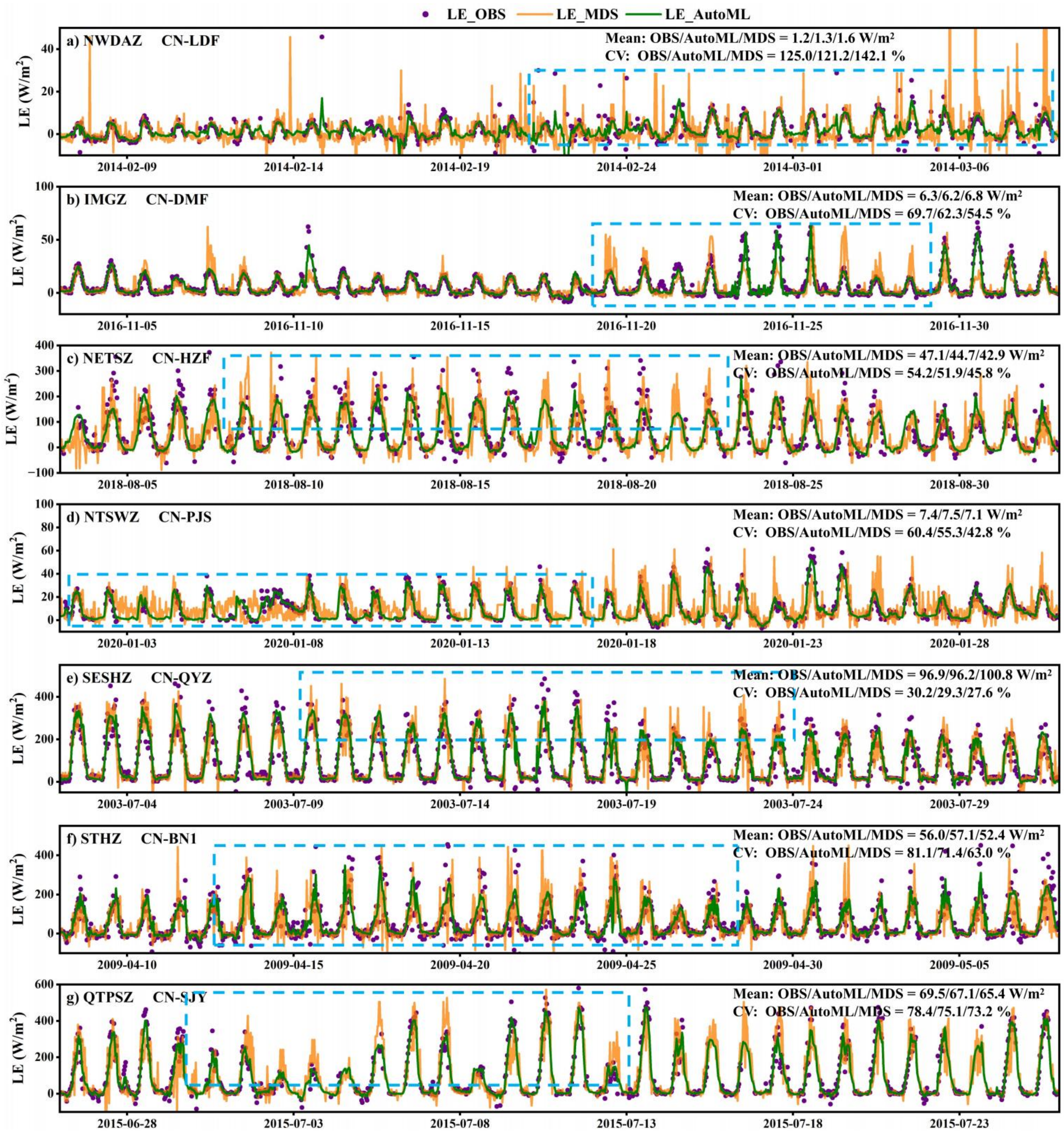


Figure 5: Comparison of gap-filled half-hourly LE time series produced by the AutoML and MDS methods across different underlying surface types under the artificial 30 day gap scenario. Blue dashed boxes highlight time periods where noticeable deviations between the MDS results and the reference observations are visually apparent, serving as illustrative examples rather than results of a formal statistical criterion.



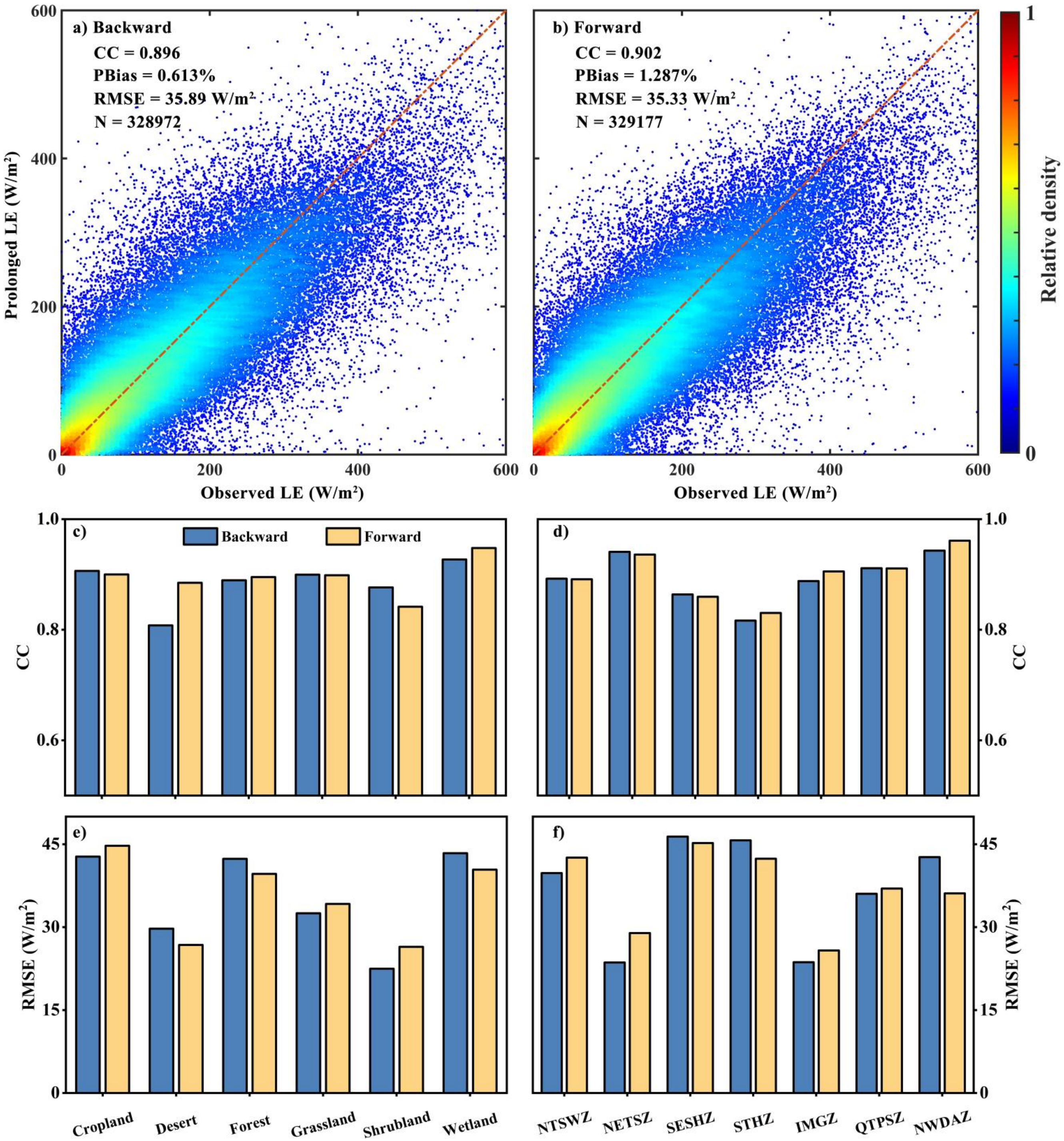
540 **Figure 6: Comparison of gap-filled half-hourly LE time series produced by the AutoML and MDS methods across different climate zones under the artificial 30 day gap scenario. Blue dashed boxes highlight time periods where noticeable deviations between the MDS results and the reference observations are visually apparent, serving as illustrative examples rather than results of a formal statistical criterion.**

3.2 Evaluation of half-hourly prolonged LE

3.2.1 Consistency between forward and backward prolongation

The prolongation results in the two temporal directions are highly consistent in a statistical sense (Fig. 7), indicating that the proposed approach is largely insensitive to the direction of temporal extrapolation and exhibits good stability. The overall scatter distributions show that both forward and backward prolongation results agree well with observations. The backward prolongation yields a correlation coefficient (CC) of 0.896 and an RMSE of 35.89 W m^{-2}, while the corresponding values for forward prolongation are 0.902 and 35.33 W m^{-2}, respectively. The differences between the two directions are small, and their overall accuracies are at comparable levels.

Across different underlying surface types, cropland, forest, grassland, and wetland exhibit high correlations under both forward and backward prolongation, with CC values generally close to or exceeding 0.89 and only minor directional differences. Among these, wetland shows high CC values in both directions but relatively larger RMSE, reflecting the larger amplitude of LE variability for this underlying surface type. In contrast, desert and shrubland display slightly lower prolongation accuracy and somewhat larger differences between forward and backward results, suggesting that under land cover conditions characterized by limited observations or higher variability, the influence of extrapolation direction becomes relatively more pronounced. Results across different climate zones further support these findings. NETSZ, IMSZ, QTPSZ, and NWDAZ maintain high consistency between forward and backward prolongation, with most CC values exceeding 0.90. Although SESHZ and STHZ exhibit relatively higher RMSE, the differences between the two temporal directions remain limited, indicating that the associated uncertainty primarily arises from climatic conditions rather than the direction of temporal prolongation. Overall, forward and backward temporal prolongation demonstrate a high degree of consistency in terms of overall accuracy and performance across underlying surface types and climate zones, with only minor differences observed under a limited number of high-uncertainty scenarios.



565 **Figure 7: Consistency between backward (a) and forward (b) prolongation of half-hourly latent heat flux (LE), shown as scatter density plots of prolonged LE versus observed LE. Each point represents one half-hourly LE record from the testing dataset used**

in the prolongation validation, and N denotes the total number of testing records included in each panel. The color scale represents the relative density of data points, normalized to highlight areas with higher concentration (warmer colors) and lower concentration (cooler colors). Panels (c) and (e) show the correlation coefficient (CC) and root mean square error (RMSE) across different underlying surface types, respectively, while panels (d) and (f) present the corresponding metrics across climate zones.

3.2.2 Temporal stability of the prolongation

To evaluate the temporal stability of the prolongation results during long-term extrapolation, models were trained using either the first 2 years or the first 6 years of available observations, and their performance was subsequently assessed on a year-by-year basis for the following periods (Fig. 8). The 2-year training scenario represents an extreme case with minimal training samples, whereas the 6-year training scenario represents more typical conditions and weather regimes at most sites. Overall, prolongation results under both training scenarios exhibit good temporal stability, with no evidence of systematic performance degradation as the extrapolation period increases. In terms of correlation, models trained on the first 2 years maintain relatively stable CC values in the range of approximately 0.88–0.93 across successive extrapolation years. Models trained on the first 6 years achieve slightly higher overall CC values and display reduced interannual variability, indicating that a larger training sample improves the consistency of prolongation results.

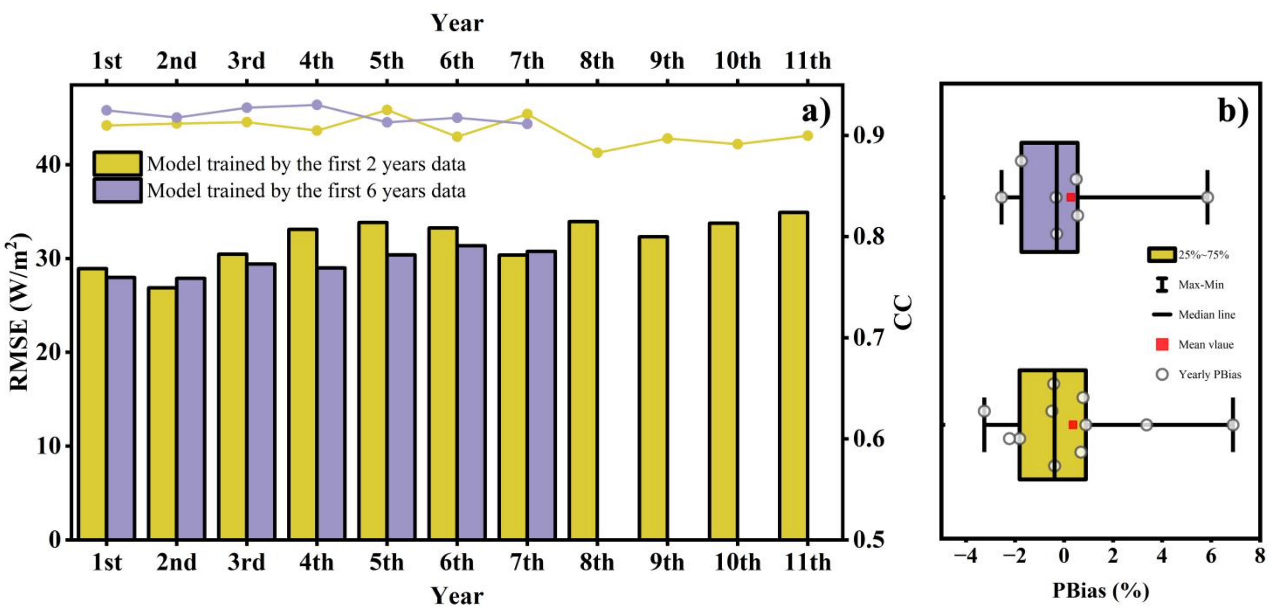


Figure 8: Temporal stability of forward prolongation performance for half-hourly LE using models trained with different lengths of observational data. In panel (a), bars represent RMSE (left y-axis), while lines represent correlation coefficient (CC, right y-axis). The x-axis "Year" indicates the calendar year of testing data outside the training interval: for sites with six or more years of observations, the first six years were used for training and subsequent years (seventh year onward) for testing; for sites with two or more years, the first two years were used for training and subsequent years for testing. Each bar and line represents the performance of the model on that year of testing data.

Error metrics show a similarly stable behavior. RMSE varies only modestly with extrapolation year, and models trained with 6 years of data generally produce lower RMSE values with smoother interannual variations. Bias analysis further indicates that PBias values under both training scenarios fluctuate around zero, with most years remaining within ±3 % and no systematic offset observed. In comparison, models trained on longer datasets exhibit more concentrated bias distributions. Taken together, these results demonstrate that the proposed temporal prolongation approach maintains high stability during long-term extrapolation. Even under the extreme scenario with limited training data, the method preserves reasonable correlations and error levels. As the length of the training dataset increases, interannual variability in CC, RMSE, and PBias is further reduced, confirming the suitability of the framework for constructing long-term continuous half-hourly LE time series.

3.3 Demonstration of different scale prolonged time series

The prolonged series at the half-hourly scale consistently reproduce the observed diurnal cycles and temporal variability (Fig. 9), as reflected by high correlation coefficients (CC ranging from 0.737 to 0.982) and generally low RMSE values (approximately 4–45 W/m^2), with PBIAS mostly within ±5% for the majority of sites. At cropland and grassland sites, the agreement is particularly strong (e.g., CC ≈ 0.98), with accurate representation of peak timing, diurnal amplitude, and overall magnitude, indicating that the dominant diurnal variability is well reproduced. Forest and wetland sites exhibit larger amplitudes and more complex short-term fluctuations, yet the prolonged series still track the dominant variability well, with deviations mainly limited to a few extreme peaks, as reflected by moderate increases in RMSE. In contrast, desert and shrubland sites show lower LE magnitudes and more irregular variability; under these conditions, the model tends to smooth isolated extreme values, resulting in slightly reduced CC or increased bias in some cases, but the overall temporal structure and diurnal patterns remain well preserved. Across climate zones, similar behavior is observed, with relatively higher uncertainty in arid and high-altitude regions, yet without introducing nonphysical discontinuities or structural distortions. These results indicate that the proposed framework can reasonably reproduce the major temporal variability and diurnal-cycle characteristics of LE at the half-hourly scale, providing support for subsequent temporal aggregation.

For the daily scale (Fig. A1) and monthly scale (Fig. A2) analyses, only periods with less than 10 % missing data at the corresponding aggregation level were considered as valid observations. Half-hourly LE records were aggregated to daily values, and daily values were further aggregated to monthly values. The AutoML-based prolonged LE data were compared with observation-based aggregated LE at representative sites. These examples include sites with varying observation lengths (from 2 to 10 years) and different climate zones and underlying surface types, aiming to demonstrate the robustness and generalization ability of the proposed method under diverse temporal and environmental conditions. Overall, the prolonged LE data consistently reproduce the seasonal variation patterns, intra-annual amplitude, and interannual variability of the observed LE at both daily and monthly scales. Compared with the half-hourly scale, temporal aggregation effectively smooths high-frequency noise, resulting in more continuous and stable temporal behavior. At cropland, grassland, and forest sites, the prolonged results show strong agreement with observations in terms of peak timing, seasonal amplitude, and long-term variability, indicating consistent agreement between prolonged and observed LE across temporal aggregation scales. For shrubland and wetland sites, where variability is higher and observational samples are relatively limited, some dispersion remains; however, the prolonged results still follow the main temporal patterns without evident systematic bias. Across different climate zones, both daily and monthly LE data exhibit good consistency with observations, indicating consistent agreement between prolonged and observed LE across temporal aggregation scales.

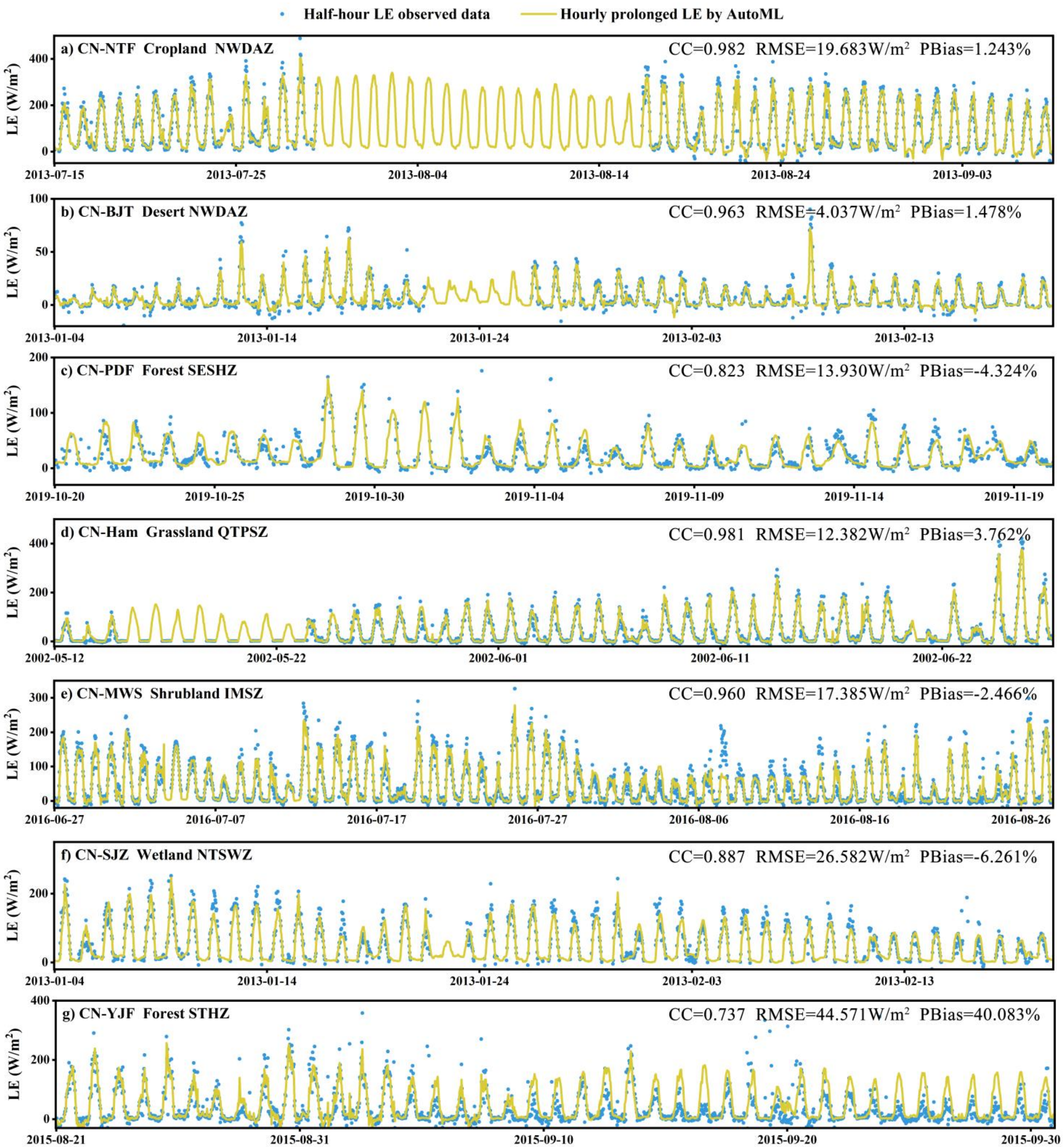


Figure 9: Demonstration of half-hourly prolonged LE time series across different typical sites.

4 Discussion

4.1 Comparison between ChinaFlux and our dataset

The preceding sections have demonstrated the stability of the proposed gap-filling and temporal prolongation framework across multiple temporal scales. To further assess the reliability of the final dataset, we compared the constructed LE dataset with the original eddy covariance observations from the 50 ChinaFlux sites used in this study (see Sect. 2.1). Given the widespread occurrence of missing data in flux measurements, only observations with zero missing rates at the corresponding temporal scale were selected as reference data for comparison. As shown in Fig. 10, the two datasets exhibit similar distributional characteristics at both the half-hourly and daily scales, considering the overall samples as well as stratifications by underlying surface type and climate zone. For most categories, the median values and interquartile ranges (25th–75th percentiles) of the reconstructed dataset closely overlap with those of the ChinaFlux observations, and the overall spread represented by the whiskers is also similar. These results indicate that the gap-filling and temporal prolongation procedures generally preserve the central tendency and variability of the original observations. Differences between the two datasets are more noticeable for desert and shrubland sites, where both datasets exhibit wider distribution ranges and larger interquartile spreads than other land-cover types, reflecting greater variability in LE under these conditions. Nevertheless, no systematic shift toward higher or lower LE values is evident in the reconstructed dataset across the examined underlying surface types or climate zones.

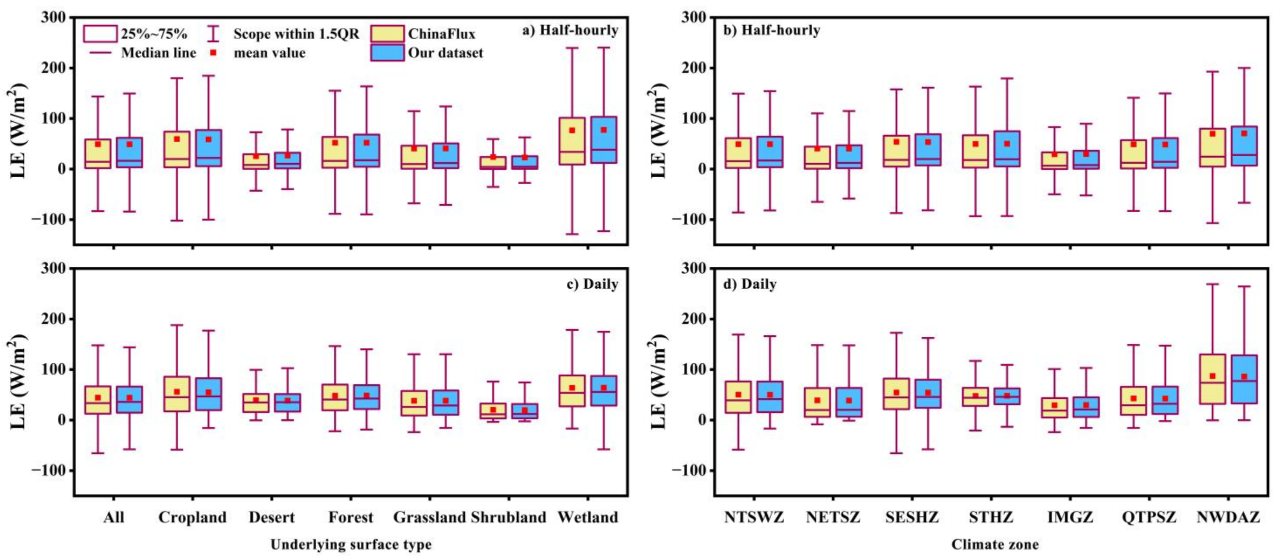


Figure 10: Comparison of data distributions between the reconstructed LE dataset and ChinaFlux observations at half-hourly (a,b) and daily (c,d) scales across different underlying surface types and climate zones.

Figure 11 presents comparisons aggregated by underlying surface type and climate zone at the daily, monthly, and annual scales. As the temporal scale progresses from daily to monthly and annual, the agreement between the two datasets further improves and dispersion decreases markedly, which is expected due to temporal aggregation and should not be interpreted as evidence of improved model performance. Instead, the key result is that the reconstructed dataset remains closely aligned with the ChinaFlux observations across aggregation scales in terms of median values, distribution ranges, and temporal variability. In particular, at the monthly and annual scales, the two datasets remain highly consistent in terms of variation magnitude and interannual trends, indicating that the prolongation procedure preserves the large-scale temporal characteristics represented by the original observations without introducing substantial systematic deviations. It should be noted that the aggregated ChinaFlux results still exhibit some dispersion at specific sites and periods, primarily due to continuous data gaps within certain months or years. In contrast, the dataset constructed in this study incorporates continuous reanalysis and remote sensing information to estimate LE at each time step during missing periods, thereby maintaining temporal continuity under conditions of limited observations. This continuous reconstruction enables complete monthly and annual aggregation even during periods affected by observational gaps, thereby improving temporal completeness while maintaining consistency with the available observations. As a result, the constructed dataset exhibits more stable statistical characteristics at the monthly and annual scales.

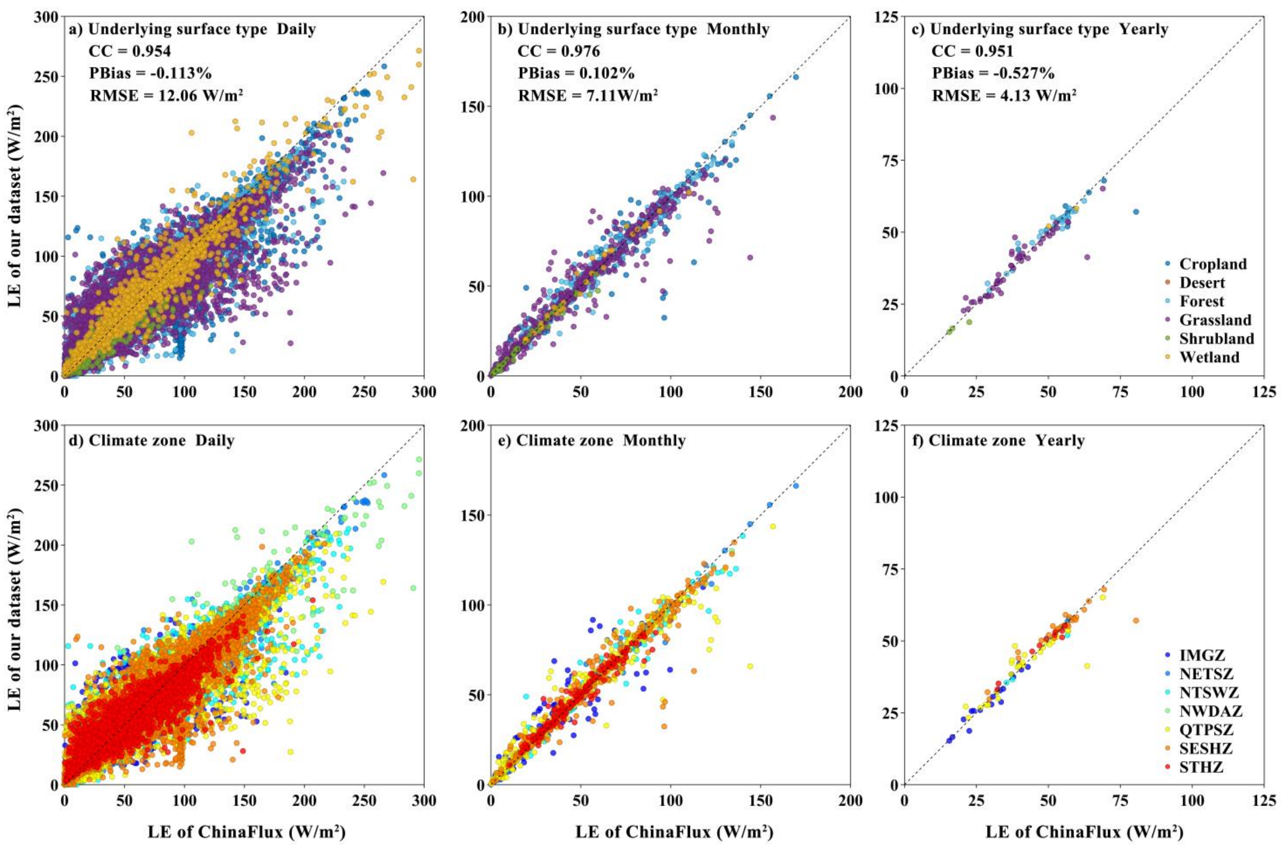


Figure 11: Comparison between the reconstructed LE dataset and ChinaFlux observations aggregated at daily (a,d), monthly (b,e), and yearly (c,f) scales for different underlying surface types and climate zones. For daily-scale comparisons, only days with less than 10 % missing half-hourly LE observations were retained as valid samples. For monthly-scale comparisons, only months with less than 10 % missing daily LE observations were included. For yearly-scale comparisons, only years with less than 10 % missing

daily LE observations were considered valid. These screening criteria were applied to ensure that the ChinaFlux reference data were primarily based on direct observations rather than gap-filled values.

4.2 Uncertainty assessment of gap-filling performance

To further evaluate the robustness of the proposed gap-filling framework, the uncertainty of AutoML-based half-hourly latent heat flux (LE) reconstruction was assessed using 20 repeated experiments under each artificial gap scenario, with different random data splits and gap realizations. The resulting distributions of CC, RMSE, and PBias are summarized in Fig. 12, providing an empirical estimate of the uncertainty associated with the gap-filling procedure. Overall, the dispersion of the three metrics remains limited across all scenarios, indicating that the model performance is not strongly sensitive to random sampling effects and that the proposed framework is statistically stable. At the shortest gap scale (30 min), the model shows the highest accuracy and the lowest uncertainty, with CC consistently concentrated around 0.91, RMSE around 30–31 W/m^2, and PBias remaining close to zero, generally within about 1 %. As gap length increases, uncertainty gradually rises, as reflected by the broader interquartile ranges and whisker spans of all three metrics. For the 1 day and 7 day scenarios, CC decreases to approximately 0.88 and 0.85, respectively, while RMSE increases to about 33–34 and 35–37 W/m^2, and PBias remains centered near zero with only moderate spread. Even under the most challenging 30 day continuous-gap scenario, the model still maintains a relatively concentrated performance distribution, with CC generally around 0.81 – 0.83 and RMSE around 36 – 39 W/m^2, although PBias exhibits a larger positive spread, with median values around 1.5 %. These results indicate that the main source of increasing uncertainty is the reduced ability to recover fine-scale LE variability under long continuous gaps, rather than the emergence of strong systematic bias. From a practical perspective, this uncertainty analysis demonstrates that the AutoML framework remains robust across repeated realizations and varying gap conditions, while also quantifying the expected decline in reconstruction confidence as gap length increases. Therefore, the repeated-experiment distributions shown in Fig. 12 provide an empirical uncertainty bound for the use of gap-filled LE data, which is particularly informative for applications involving long missing periods or uncertainty-sensitive analyses.

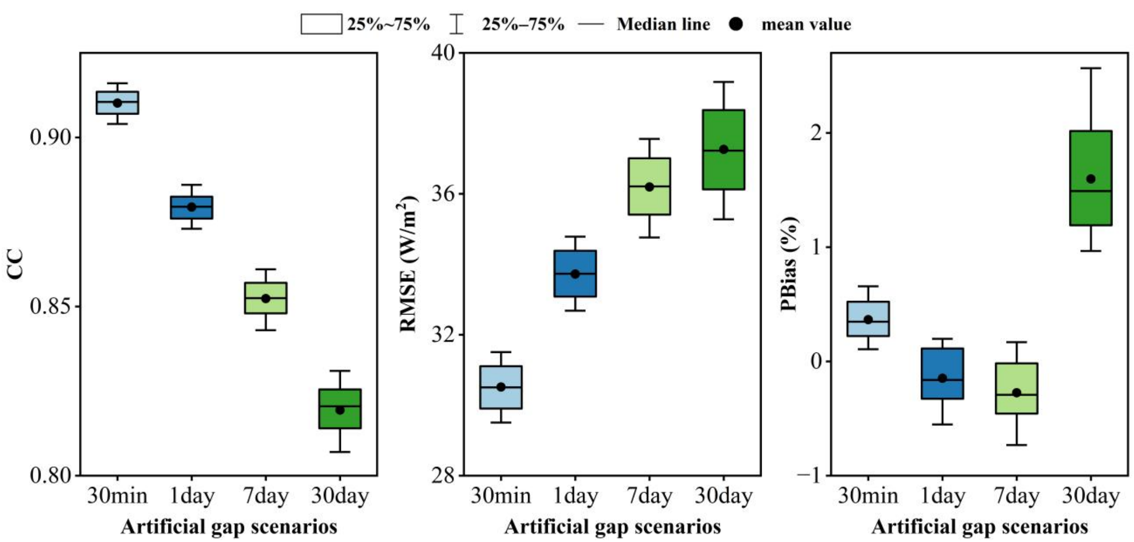


Figure 12: Uncertainty of AutoML-based half-hourly LE gap-filling performance under different artificial gap scenarios (30 min, 1 day, 7 day, and 30 day), estimated from 20 repeated experiments with different random data splits and gap realizations. Boxplots show the distributions of correlation coefficient (CC), root mean square error (RMSE), and percentage bias (PBias).

4.3 Importance of input features

To further understand how the model represents latent heat flux (LE) under different environmental conditions, the SHapley Additive exPlanations (SHAP) method was applied to quantify the relative importance of input features and their contributions to model predictions. Figures 13 and 14 present the SHAP results for the full sample and for grouped analyses based on underlying surface types and climate zones, respectively, illustrating the relative contributions and distributions of individual predictors to LE estimation. For the full sample (Fig. 13a), variables directly related to surface energy supply dominate the model, with net radiation (Rn) and energy-related terms associated with LE exhibiting the highest importance. This result indicates that the model broadly adheres to the fundamental physical constraints of evapotranspiration, whereby energy availability is the primary control on LE variability. At the same time, vegetation-related variables such as leaf area index (LAI) and normalized difference vegetation index (NDVI) also show substantial importance, highlighting the significant regulatory role of vegetation growth and seasonality in evapotranspiration processes. In contrast, variables characterizing atmospheric conditions and water availability (e.g., vapor pressure deficit (VPD), soil moisture, and precipitation) contribute less at the full-sample level, although they still play important complementary roles under specific environmental conditions.

Across different underlying surface types (Fig. 13b–g), the ranking of feature importance exhibits pronounced differences. For cropland, forest, and grassland, energy-related variables and vegetation indices jointly dominate LE variability, indicating that in ecosystems with relatively dense vegetation cover, evapotranspiration is simultaneously controlled by radiative conditions and vegetation physiological activity. In contrast, for desert and shrubland, the importance of vegetation-related variables decreases markedly, while radiation- and atmosphere-related factors become relatively more influential, suggesting that under sparse vegetation conditions, LE is more directly constrained by energy input and atmospheric evaporative demand. This is also consistent with the dominance of evaporation over transpiration in these ecosystems, where limited vegetation cover reduces the contribution of plant physiological processes, leading to a weaker explanatory power of vegetation-related variables. Wetlands show relatively high contributions from both energy-related and water-related variables, reflecting the critical role of water availability in regulating evapotranspiration in these environments.

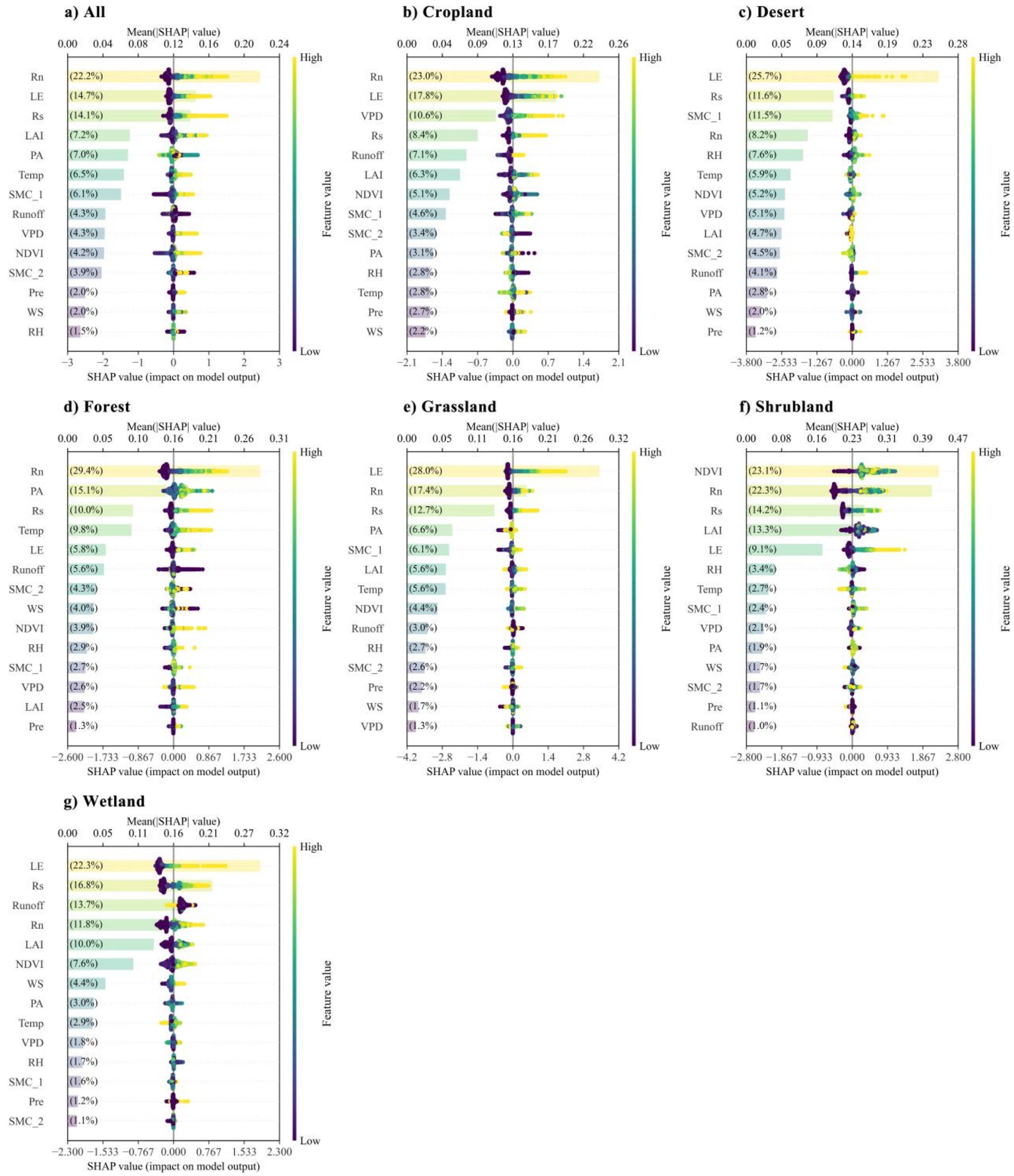


Figure 13: SHAP-based feature importance and contribution patterns for LE prediction across all samples and different underlying surface types. Bars represent the mean absolute SHAP values, which quantify the average contribution of each predictor to model output and are shown together with their relative importance percentages. In the beeswarm plots, each point represents one half-hourly sample, and its horizontal position indicates the SHAP value of the corresponding predictor. Positive SHAP values indicate a positive contribution to predicted LE, whereas negative values indicate a negative contribution. Point colors represent the normalized magnitude of the predictor value, ranging from low (purple) to high (yellow).

The SHAP results across different climate zones (Fig. 14) further elucidate how the model responds to regional climatic differences. Overall, energy-related variables consistently dominate across all climate zones, but their relative rankings and secondary controlling factors vary substantially among regions. In humid and subhumid climate zones (e.g., NTSWZ, NETSZ, and SESHZ), net radiation (Rn) and shortwave radiation (Rs) consistently rank among the most important predictors, indicating that LE variability is primarily controlled by surface energy supply. Meanwhile, LE itself and vegetation state variables such as LAI and NDVI contribute persistently, albeit to a lesser extent than energy terms, reflecting the modulation of evapotranspiration by vegetation seasonality. By contrast, atmospheric and soil moisture variables (e.g., VPD, soil moisture content (SMC), and precipitation) generally exhibit lower importance and mainly act as secondary regulators. In arid and semiarid climate zones (e.g., IMSZ, QTPSZ, and NWDAZ), the relative importance of energy-related variables (Rn and Rs) and LE further increases, while the contribution of vegetation indices declines markedly. Sparse vegetation cover in these regions results in increased exposure of bare soil surfaces, leading to a greater contribution of soil evaporation relative to plant transpiration and thereby reducing the explanatory power of vegetation-related variables. At the same time, atmospheric variables such as air temperature and VPD enter the upper ranks of importance in some regions, indicating that under water-limited conditions, LE is more directly constrained by the combined effects of energy input and atmospheric evaporative demand. In particular, in the plateau climate zone (QTPSZ), air temperature and soil moisture variables also exhibit non-negligible importance in addition to radiative factors, reflecting the integrated modulation of evapotranspiration by complex terrain and low-temperature environments.

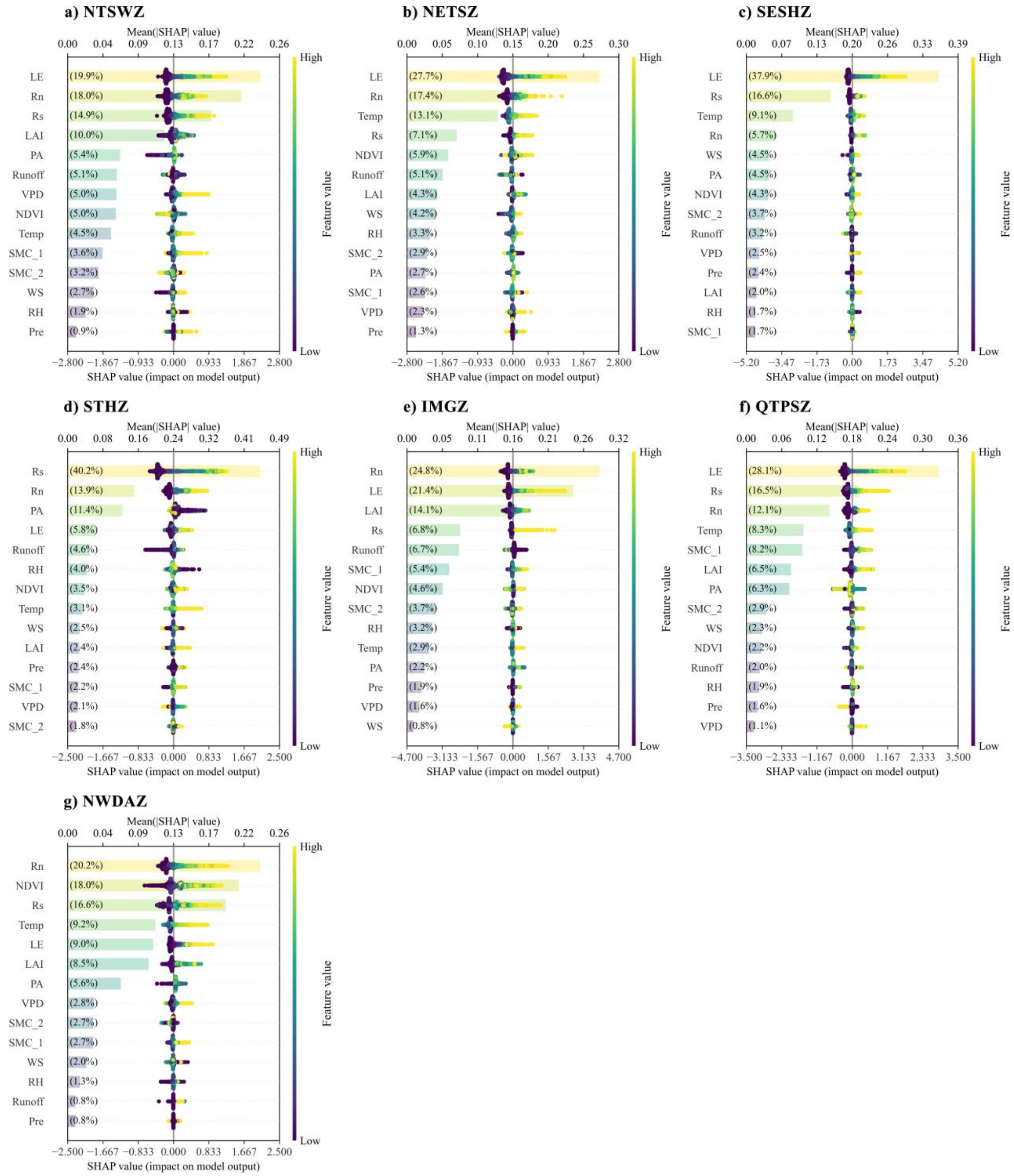


Figure 14: SHAP-based feature importance and contribution patterns for LE prediction across different climate zones. Bars represent the mean absolute SHAP values, which quantify the average contribution of each predictor to model output and are shown together with their relative importance percentages. In the beeswarm plots, each point represents one half-hourly sample, and its horizontal position indicates the SHAP value of the corresponding predictor. Positive SHAP values indicate a positive contribution to predicted LE, whereas negative values indicate a negative contribution. Point colors represent the normalized magnitude of the predictor value, ranging from low (purple) to high (yellow).

5 Data availability

This study publicly releases a temporally continuous half-hourly latent heat flux (LE) dataset for flux tower sites across China, together with multi-temporal-scale aggregated products and key driving variables used for gap-filling and temporal prolongation. The dataset covers the period from 2000 to 2024 and includes the following data products.

1) Half-hourly LE data (2000–2024). The half-hourly LE dataset represents the core product of this study and provides complete time series for all flux tower sites. Each record is accompanied by a quality flag (QC_LE) indicating its data source: T denotes half-hourly LE derived from original flux tower observations; F represents gap-filled LE generated within the observation period using the AutoML framework; and P indicates LE values produced through temporal prolongation beyond the observation period using the same framework. For the 40 sites with original observations, T corresponds to original QA/QC-filtered tower observations. For the 10 marked sites, however, T represents provider-supplied complete LE time series within the nominal observation period and should not be interpreted as purely original observations. This distinction is explicitly stated in Appendix Table A1 and in the Readme.txt file. All half-hourly data follow a unified time format, with timestamps referenced to local site time to ensure temporal continuity and traceability. Timestamps are provided in the format *YYYYMMDD–HHMM* (e.g., 20120101–1530), representing local date and time at each site. Each timestamp denotes the beginning of the corresponding 30 min flux-averaging period. For example, the records for 1 January 2012 range from 20120101–0000 to 20120101–2330, representing 48 half-hourly time steps, and 20120101–1530 indicates the start of the 15:30–16:00 local-time interval.

2) Daily, monthly, and annual aggregated data. Daily, monthly, and annual LE products were directly aggregated from the gap-filled and temporally prolonged half-hourly LE time series and cover the period 2000–2024. All aggregated products originate from the same half-hourly dataset, ensuring full temporal consistency across different time scales. For each temporal scale, instead of assigning a single categorical quality flag, we provide N_obs, defined as the number of half-hourly observations (QC_LE = T) contributing to the aggregated value (e.g., 0–48 for daily data). This explicit representation of observation availability allows users to assess the proportion of measured versus reconstructed data within each aggregated value, enabling flexible data screening based on research needs while retaining stable and continuous datasets for long-term hydrological and climate change studies.

3) Auxiliary driving data. To enhance reproducibility and extensibility, key driving variables used for LE gap-filling and temporal prolongation are also provided. These include hourly meteorological and energy balance–related variables from ERA5-Land, as well as NDVI and LAI products from MODIS. All auxiliary datasets have been spatially and temporally matched to the LE data and can be directly used for method reproduction or extended analyses.

In addition, the code used for LE gap-filling and temporal prolongation is publicly available and provided alongside the dataset, enabling full reproducibility of the data processing workflow and facilitating reuse for similar applications.

All data products are released in CSV format, with file names clearly indicating temporal scale and variable type to facilitate data access and use. The product has been deposited at https://doi.org/10.5281/zenodo.18194590 (Qian et al., 2026) and can be downloaded publicly.

6 Conclusion

This study developed a unified gap-filling and temporal prolongation framework for half-hourly latent heat flux (LE) observations and produced a continuous benchmark dataset based on 50 ChinaFlux sites across China. By integrating eddy covariance observations with remote sensing and reanalysis information, the dataset substantially reduces the impact of data gaps and heterogeneous observation periods that commonly limit the direct use of flux tower measurements.

The primary contribution of this work is the establishment of a long-term, half-hourly LE benchmark dataset covering the period 2000–2024. Compared with existing site observations, the reconstructed dataset provides temporally continuous records while preserving the major temporal characteristics and statistical properties of the original measurements across a wide range of land-cover types and climate zones. This improves the usability of ChinaFlux observations for applications that require continuous long-term records, including evapotranspiration product evaluation, land surface model benchmarking, water–energy cycle analysis, and climate change studies.

In addition to the LE dataset itself, the study provides a standardized framework for addressing missing observations and unequal observation periods in flux networks. The methodology is designed to maximize the value of existing eddy covariance measurements while maintaining consistency with the physical controls governing evapotranspiration processes. The resulting benchmark dataset offers an observation-based reference for evaluating and improving remote sensing products and model simulations across diverse environmental conditions in China.

By providing continuous half-hourly LE records together with supporting auxiliary data and source information, this dataset enhances the accessibility and long-term usability of ChinaFlux observations. We anticipate that it will serve as a valuable resource for the broader hydrological, ecological, and climate research communities and facilitate future studies of terrestrial water, energy, and carbon interactions across China.

Appendix A

Table A1. Basic information for 50 sites in China. Stations marked with "*" provide only interpolated time series data (without QA/QC-filtered observations). Elevation refers to the ground elevation of the flux tower site above mean sea level. Height refers to the heights of the towers above ground (the value in parentheses). The Number of LE Data and Missing Ratio columns are calculated from these quality-screened reliable records. The forest sites are subdivided as follows, CN-BN1: Tropical Seasonal Rainforest; CN-BTM: Natural Oak Forest; CN-CBS: Broad-Leaved Red Pine Forest; CN-DHS: Mixed Coniferous Broad-Leaved Forest; CN-HHL: Poplar−Tamarisk Mixed Forest; CN-HYL: Euphrates Poplar Forest; CN-HZF: Temperate Boreal Coniferous Forest; CN-JFS: Subtropical Evergreen Broad-Leaved Forest; CN-PDF: Shrub-Dominated Secondary Forest; CN-QYZ: Coniferous Plantation; CN-XLD: Quercus variabilis Plantation; CN-YS3: Coniferous Plantation.

Station ID	SITE Name	Longitude	Latitude	Climate zone	Underlying surface type	Elevation and Height	Annual average temperature	Annual precipitation	Number of LE Data	Missing ratio	Start yea	End year
		(°E)	(°N)			(m)	(°C)	(mm)				
1*	CN-JFS	107.1508	29.0217	SESHZ	Forest	1543 (28)	8.2	1434	35088	0.000	2020	2021
2*	CN-NQF	92.0167	31.6500	QTPSZ	Grassland	4585 (2.3)	1.9	430	87648	0.000	2014	2018
3*	CN-QYZ	115.0667	26.7333	SESHZ	Forest	110 (39.6)	17.9	1489	227904	0.000	2003	2015
4*	CN-XLD	112.4667	35.0167	NTSWZ	Forest	410 (36)	13.4	641.7	35088	0.000	2016	2017
5*	CN-DHS	112.5343	23.1738	STHZ	Forest	300 (38)	22.5	1714	140256	0.000	2003	2011
6*	CN-HB3	101.3333	37.6667	QTPSZ	Grassland	3200 (2.5)	-1.7	580	140256	0.000	2003	2011
7*	CN-DXF	91.0833	30.8500	QTPSZ	Grassland	4333 (2.2)	1.3	477	122736	0.000	2004	2011
8*	CN-NMG	116.4040	43.3255	IMGZ	Grassland	1317 (2.5)	2.1	365	122736	0.000	2004	2011
9	CN-DSL	98.9406	38.8399	QTPSZ	Wetland	3439 (4.5)	-1.2	410	44427	0.001	2014	2016
10*	CN-YJF	102.1775	23.4739	STHZ	Grassland	553 (13.9)	23.8	711.8	46800	0.001	2013	2016
11*	CN-BTM	111.9352	33.4997	SESHZ	Forest	1410.7 (38)	15.2	830	35040	0.001	2017	2019
12	CN-CLF	123.4703	44.5966	NETSZ	Cropland	143 (2)	6.4	500	43617	0.009	2018	2021
13	CN-Cng	123.5092	44.5934	NETSZ	Grassland	145 (2)	5.6	420	58312	0.023	2007	2010
14	CN-HaM	101.1800	37.3700	QTPSZ	Grassland	3150 (2.2)	0.8	520	50439	0.036	2002	2004
15	CN-MWS	107.2300	37.7100	IMSZ	Shrubland	1530 (6.2)	8.3	293	83863	0.044	2012	2016
16	CN-HB1	101.3167	37.6167	QTPSZ	Grassland	3200 (2.5)	-1.7	580	100449	0.045	2004	2009
17	CN-LCA	114.6833	37.8833	NTSWZ	Cropland	50.1 (3.5)	13.4	341	66910	0.046	2013	2017
18	CN-XLH	116.6714	43.5544	IMGZ	Grassland	1250 (4)	2.6	349	162766	0.053	2006	2015
19	CN-JRF	119.2173	31.8068	SESHZ	Cropland	15 (3.5)	15.7	1080	98580	0.059	2015	2020
20	CN-GCF	115.6667	39.1333	NTSWZ	Cropland	15 (4.5)	12.2	528	43351	0.136	2020	2022
21	CN-YCA	116.5702	36.8290	NTSWZ	Cropland	28 (3.3)	13.1	528	120296	0.143	2003	2011
22	CN-REG	102.5500	32.8000	QTPSZ	Grassland	3500 (2.5)	1.5	747	83600	0.153	2015	2020
23	CN-DTH	113.0525	29.4875	SESHZ	Wetland	68 (6.5)	16.4	1382	39638	0.197	2006	2008
24	CN-BN1	101.2653	21.9275	STHZ	Forest	756 (72)	21.5	1556.8	108316	0.228	2003	2011
25	CN-CBS	128.0958	42.4025	NETSZ	Forest	1080 (61.5)	3.8	800	99328	0.278	2003	2010
26	CN-JZF	121.2017	41.1480	NTSWZ	Cropland	23 (5)	8.5	640	110654	0.369	2005	2014
27	CN-PJS	121.9646	40.9328	NTSWZ	Wetland	3.8 (4.2)	9.5	631	31866	0.393	2018	2020

28	CN-SJZ	118.9809	37.7664	NTSWZ	Wetland	3.5 (3)	12.66	604	84696	0.396	2011	2018
29	CN-Du2	116.2836	42.0466	IMGZ	Grassland	1350 (5)	2.4	380	39647	0.413	2015	2018
30	CN-HZF	121.0178	51.7811	NETSZ	Forest	773 (37)	-4.4	481.6	49790	0.432	2014	2018
31	CN-Hgu	102.5900	32.8453	QTPSZ	Grassland	3480 (2.5)	1.2	720	25881	0.438	2015	2017
32	CN-PDF	106.3167	26.6000	SESHZ	Forest	1170 (24)	15.96	1432	46063	0.451	2015	2019
33	CN-LDF	101.1326	41.9993	NWDAZ	Desert	878 (3)	9.75	37.31	22885	0.473	2013	2015
34	CN-YS1	116.6563	40.4190	NTSWZ	Shrubland	141 (30)	12.5	580	18426	0.473	2020	2021
35	CN-SJY	100.4800	34.3547	QTPSZ	Grassland	3950 (2.2)	-2.9	531	45943	0.476	2012	2016
36	CN-SSW	100.4933	38.7892	NWDAZ	Desert	1594 (4.6)	7.29	184.83	20520	0.486	2013	2015
37	CN-HB2	101.3119	37.6094	QTPSZ	Grassland	3200 (2.2)	-1.7	550	53371	0.493	2015	2020
38	CN-NTF	101.1338	42.0048	NWDAZ	Cropland	875 (3.5)	9.75	37.31	20381	0.493	2013	2015
39	CN-HYL	101.1239	41.9932	NWDAZ	Forest	876 (22)	9.75	37.31	21210	0.507	2013	2015
40	CN-ARF	100.4643	38.0473	QTPSZ	Grassland	3033 (3.5)	-0.8	410	59548	0.515	2013	2019
41	CN-YS3	116.6588	40.4165	NTSWZ	Forest	328 (30)	12.5	580	16344	0.534	2020	2021
42	CN-ZYF	100.4464	38.9751	NWDAZ	Wetland	1460 (5.2)	7.29	184.83	75913	0.567	2013	2022
43	CN-BJT	100.3042	38.9150	NWDAZ	Desert	1562 (4.6)	7.29	184.83	14415	0.571	2013	2014
44	CN-DMF	110.3315	41.6439	IMGZ	Grassland	1409 (2.3)	4.6	255.2	69214	0.605	2013	2022
45	CN-HZZ	100.3186	38.7652	NWDAZ	Desert	1731 (4.6)	7.29	184.83	66672	0.620	2013	2022
46	CN-YKF	100.2421	38.0142	QTPSZ	Grassland	4146 (3.2)	2.3	41.0	26064	0.628	2015	2018
47	CN-HHL	101.1335	41.9903	NWDAZ	Forest	874 (22)	9.75	37.31	58114	0.647	2013	2022
48	CN-DMZ	100.3722	38.8555	NWDAZ	Cropland	1556 (4.5)	7.29	184.83	38125	0.660	2013	2019
49	CN-HMF	100.9872	42.1135	NWDAZ	Desert	1054 (4.5)	9.75	37.31	34189	0.746	2015	2022
50	CN-QJF	118.2500	31.9667	SESHZ	Cropland	18 (1.5)	15.8	1090	12543	0.762	2017	2020

Table A2. Data Source Information for flux towers.

Station_ID	Reference DOI	Data DOI
1	10.11922/11-6035.csd.2023.0103.zh	http:/dx.doi.org/10.57760/sciencedb.o00119.00048
2	10.11922/11-6035.csd.2024.0098.zh	https://doi.org/10.57760/sciencedb.j00001.00821
3	10.11922/11-6035.csd.2024.0142.zh	https://doi.org/10.57760/sciencedb.ecodb.00105
4	10.11922/11-6035.csd.2023.0082.zh	https://doi.org/10.57760/sciencedb.ecodb.00095
5	10.11922/csdata.2020.0046.zh	http://www.dx.doi.org/10.11922/sciencedb.1009
6	10.11922/csdata.2020.0034.zh	http://www.sciencedb.cn/dataSet/handle/1007
7	10.11922/csdata.2020.0026.zh	http://www.dx.doi.org/10.11922/sciencedb.j00001.00248
8	https://doi.org/10.12199/nesdc.ecodb.chinaflux2003-2010.2021.nmg.006	https://doi.org/10.12199/nesdc.ecodb.chinaflux2003-2010.2021.nmg.006
9	10.11922/11-6035.csd.2024.0141.zh	https://www.scidb.cn/doi/10.11888/Atmos.tpdc.301233
10	10.11922/csdata.2020.0057.zh	http://www.dx.doi.org/10.11922/sciencedb.j00001.00066
11	10.11922/11-6035.csd.2023.0124.zh	http://www.dx.doi.org/10.57760/sciencedb.j00001.00859
12	10.11922/11-6035.csd.2023.0048.zh	https:/doi.org/10.57760/sciencedb.o00119.00068
13	https://doi.org/10.18140/FLX/1440209	https://fluxnet.org/sites/siteinfo/CN-Cng
14	https://doi.org/10.18140/FLX/1440190	https://fluxnet.org/sites/siteinfo/CN-HaM
15	10.17521/cjpe.2023.0001	https://doi.org/10.57760/sciencedb.ecodb.00094
16	10.11922/csdata.2020.0033.zh	http://www.sciencedb.cn/dataSet/handle/1010
17	10.11922/11-6035.csd.2023.0031.zh	https://doi.org/10.57760/sciencedb.o00119.00050
18	10.11922/11-6035.csd.2023.0059.zh	http://doi.org/10.57760/sciencedb.07140
19	10.11922/11-6035.csd.2023.0072.zh	http://dx.doi.org/10.57760/sciencedb.07131
20	10.11922/11-6035.csd.2023.0022.zh	https://doi.org/10.57760/sciencedb.o00119.00071
21	10.11922/csdata.2020.0044.zh	http://www.doi.org/10.11922/sciencedb.j00001.20002
22	10.11922/11-6035.csd.2023.0009.zh	https://dx.doi.org/10.57760/sciencedb.o00119.00052
23	10.11922/11-6035.csd.2023.0055.zh	https://doi.org/10.57760/sciencedb.o00119.00070
24	10.11922/csdata.2020.0037.zh	http://www.dx.doi.org/10.11922/sciencedb.j00001.00177
25	https://www.sciengine.com/CSD/doi/10.11922/csda	http://www.dx.doi.org/10.11922/sciencedb.1000

	ta.2020.0041.zh	
26	10.11922/11-6035.csd.2023.0007.zh	https://doi.org/10.57760/sciencedb.o00119.00062
27	10.11922/11-6035.csd.2023.0004.zh	https://doi.org/10.57760/sciencedb.06829
28	10.11922/11-6035.ncdc.2023.0007.zh	https://doi.org/10.57760/sciencedb.j00001.00867
29	https://doi.org/10.18140/FLX/1440140	https://fluxnet.org/sites/siteinfo/CN-Du2
30	10.11922/11-6035.csd.2023.0019.zh	https://doi.org/10.57760/sciencedb.o00119.00063
31	https://doi.org/10.18140/FLX/1669632	https://fluxnet.org/sites/siteinfo/CN-Hgu
32	10.11922/11-6035.csd.2023.0063.zh	https://doi.org/10.57760/sciencedb.o00119.00074
33	10.11922/11-6035.csd.2024.0099.zh	https://www.scidb.cn/doi/10.11888/Atmos.tpdc.301184
34	10.11922/11-6035.csd.2023.0030.zh	https://doi.org/10.57760/sciencedb.o00119.00035
35	10.11922/11-6035.csd.2023.0056.zh	https://doi.org/10.57760/sciencedb.o00119.00025
36	10.11922/11-6035.csd.2024.0099.zh	https://www.scidb.cn/doi/10.11888/Atmos.tpdc.301184
37	10.11922/11-6035.csd.2023.0012.zh	https://dx.doi.org/10.57760/sciencedb.06764
38	10.11922/11-6035.csd.2024.0099.zh	https://www.scidb.cn/doi/10.11888/Atmos.tpdc.301184
39	10.11922/11-6035.csd.2024.0099.zh	https://www.scidb.cn/doi/10.11888/Atmos.tpdc.301184
40	10.11922/11-6035.csd.2024.0141.zh	https://www.scidb.cn/doi/10.11888/Atmos.tpdc.301233
41	10.11922/11-6035.csd.2023.0030.zh	https://doi.org/10.57760/sciencedb.o00119.00035
42	10.11922/11-6035.csd.2024.0099.zh	https://www.scidb.cn/doi/10.11888/Atmos.tpdc.301184
43	10.11922/11-6035.csd.2024.0099.zh	https://www.scidb.cn/doi/10.11888/Atmos.tpdc.301184
44	10.11922/11-6035.csd.2023.0021.zh	https://doi.org/10.57760/sciencedb.o00119.00043
45	10.11922/11-6035.csd.2024.0099.zh	https://www.scidb.cn/doi/10.11888/Atmos.tpdc.301184
46	10.11922/11-6035.csd.2024.0141.zh	https://www.scidb.cn/doi/10.11888/Atmos.tpdc.301233
47	10.11922/11-6035.csd.2024.0099.zh	https://www.scidb.cn/doi/10.11888/Atmos.tpdc.301184

48	10.11922/11-6035.csd.2024.0099.zh	https://www.scidb.cn/doi/10.11888/Atmos.tpdc.301184
49	10.11922/11-6035.csd.2024.0099.zh	https://www.scidb.cn/doi/10.11888/Atmos.tpdc.301184
50	10.11922/11-6035.csd.2023.0001.zh	https://doi.org/10.57760/sciencedb.j00001.00779

815

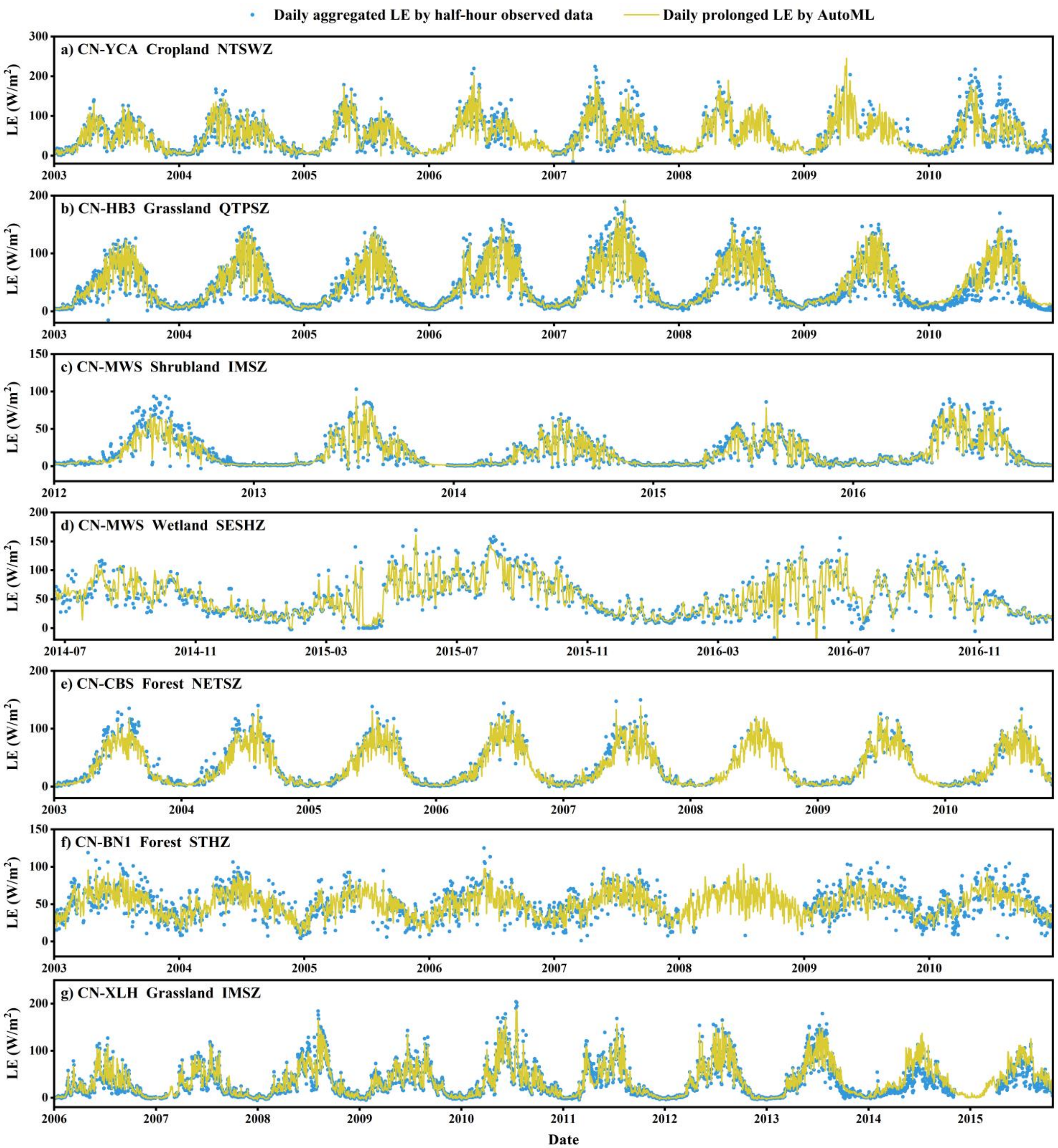


Figure A1: Demonstration of daily prolonged LE time series across different typical sites.

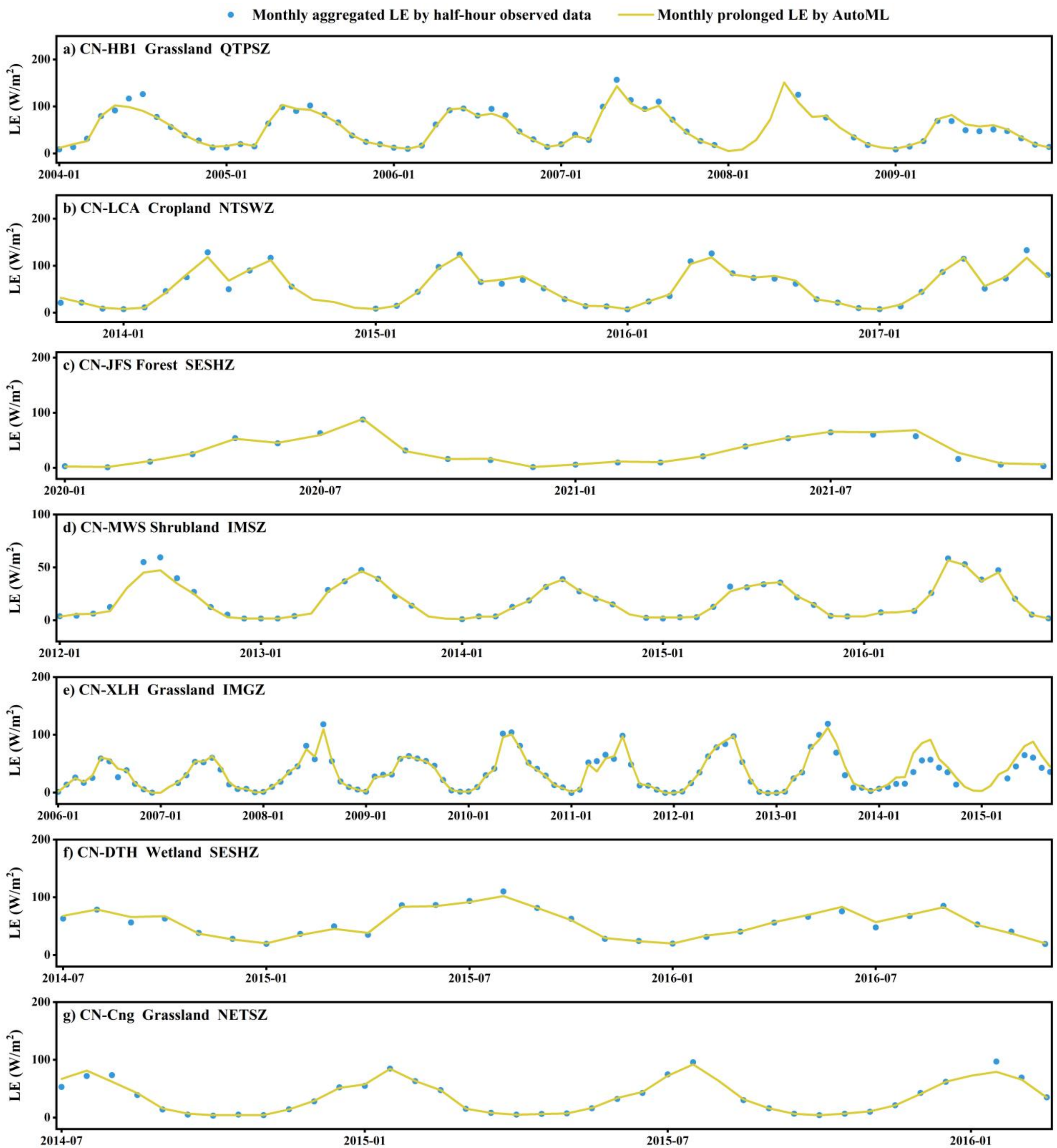


Figure A2: Demonstration of monthly prolonged LE time series across different typical sites.

Author contributions

Conceptualization: Long Qian and Lifeng Wu. Methodology: Long Qian, Lifeng Wu, and Xingjiao Yu. Data curation: Long Qian. Funding acquisition: Lifeng Wu, Zhitao Zhang, and Junying Chen. Writing (initial): Long Qian. Writing (review and editing): Xiaogang Liu, Sien Li and Sumeng Ye. Supervision: Rangjian Qiu Xuqian Bai, and Yaokui Cui.

Competing interests

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

Disclaimer

Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

Acknowledgements

The authors gratefully acknowledge the ChinaFlux research network for providing eddy covariance flux observations used in this study. The ChinaFlux data were obtained through the National Ecosystem Science Data Center (https://www.nesdc.org.cn/, last access: 11 November 2025). We thank the European Centre for Medium-Range Weather Forecasts (ECMWF) for making the ERA5-Land reanalysis data publicly available (https://www.ecmwf.int/en/era5-land, last access: 11 November 2025). We also acknowledge the MODIS science team for providing the MOD13Q1, MYD13Q1, MOD15A2H, and MYD15A2H products used to derive vegetation indices in this study. The authors thank the Google Earth Engine platform (https://earthengine.google.com/, last access: 11 November 2025) for enabling efficient access to and processing of the ERA5-Land and MODIS datasets. We further acknowledge the H2O.ai team for developing the AutoML of H2O framework (https://www.h2o.ai/, last access: 11 November 2025), which was used for gap-filling and temporal prolongation of latent heat flux time series.

Financial support

This research has been supported by the Special project of scientific and technological innovation of Xinjiang Research Institute of Arid Area Agriculture (XJHQNY-2025-3), Tianjin Key Laboratory of Rail Transit Navigation Positioning and Spatio-temporal Big Data Technology under Grant (TKL2024B12), National Key Research and Development Program of China (2023YFD1900702-01, 2022YFD1900404), the National Natural Science Foundation of China (52479050).

References

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

Allen, R. G., Pereira, L. S., Raes, D., and Smith, M.: Crop evapotranspiration – guidelines for computing crop water requirements, FAO Irrigation and Drainage Paper No. 56, FAO, Rome, Italy, 1998.

Chen, Y. Y., Chu, C. R., and Li, M. H.: A gap-filling model for eddy covariance latent heat flux: Estimating evapotranspiration of a subtropical seasonal evergreen broad-leaved forest as an example, J. Hydrol., 468–469, 101–110, https://doi.org/10.1016/j.jhydrol.2012.08.026, 2012.

Cheng, M., Jiao, X., Jin, X., Li, B., Liu, K., and Shi, L.: Satellite time series data reveal interannual and seasonal spatiotemporal evapotranspiration patterns in China in response to effect factors, Agric. Water Manag., 255, 107046, https://doi.org/10.1016/j.agwat.2021.107046, 2021.

Copernicus Climate Change Service (C3S): ERA5-Land hourly data from 1950 to present, Climate Data Store (CDS), https://doi.org/10.24381/cds.e2161bac, 2022.

Didan, K.: MOD13Q1 MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V006, NASA LP DAAC, https://doi.org/10.5067/MODIS/MOD13Q1.006, 2015.

Didan, K., and Barreto Munoz, A, MODIS Vegetation Index User's Guide (MOD13 Series), Version 3.10, Collection 6.1. University of Arizona / LP DAAC, User Guide. https://lpdaac.usgs.gov/documents/621/MOD13_User_Guide_V61.pdf. 2019.

Elnashar, A., Wang, L. L., Wu, B. F., Zhu, W. W., and Zeng, H. W.: Synthesis of global actual evapotranspiration from 1982 to 2019, Earth Syst. Sci. Data, 13, 447–480, https://doi.org/10.5194/essd-13-447-2021, 2021.

Fan, J. L., Yue, W. J., Wu, L. F., Zhang, F. C., Cai, H. J., Wang, X. K., Lu, X. H., and Xiang, Y. Z.: Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of China, Agric. For. Meteorol., 263, 225–241, https://doi.org/10.1016/j.agrformet.2018.08.019, 2018.

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 Resour. Res., 53, 2618–2626, https://doi.org/10.1002/2016WR020175, 2017.

Friedl, M.A., Sulla-Menashe, D., Tian, B., Schneider, A., Ramankutty. N., Sibley. A., Huang, X.M.: MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets, Remote Sensing of Environment, 114(1), 168-182, https://doi.org/10.1016/j.rse.2009.08.016, 2010.

Guo, D., Parehkar, A., Ryu, D., Wang, Q. J., and Western, A. W.: Parsimonious gap-filling models for sub-daily actual evapotranspiration observations from eddy-covariance systems, Remote Sens., 14, 1286, https://doi.org/10.3390/rs14051286, 2022.

Guo, L. N., Wu, Y. H., Zheng, H. X., Zhang, B., Fan, L. X., Chi, H. J., Yan, B. K., and Wang, X.Q.: Consistency and uncertainty of gridded terrestrial evapotranspiration estimations over China, J. Hydrol., 612, 128245, https://doi.org/10.1016/j.jhydrol.2022.128245, 2022.

Guo, N., Chen, H., Han, Q., and Wang, T. J.: Evaluating data-driven and hybrid modeling of terrestrial actual evapotranspiration based on an automatic machine learning approach, Journal of Hydrology, 628, 130594, https://doi.org/10.1016/j.jhydrol.2023.130594, 2024.

Han, Q. Z., Liu, L. Y., and Liu, X. J.: Upscaling Tower-Based Net Ecosystem Productivity to 250 m Resolution with Flux Site Distribution Considerations, Remote Sensing, 17(3), 426, https://doi.org/10.3390/rs17030426, 2025.

H2O.ai.: H2O AutoML Documentation, https://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html, 2023.

He, X. L., Liu, S. M., Bateni, S., Xu, T. R., Jun, C., Kim, D., Li, X., Song, L. S., Zhao, L., Xu, Z. W., and Wei, J. X.: Innovative approach for estimating evapotranspiration and gross primary productivity by integrating land data assimilation, machine learning, and multi-source observations, Agric. For. Meteorol., 355, 110136, https://doi.org/10.1016/j.agrformet.2024.110136, 2025.

Jung, M., Koirala, S., Weber, U., Ichii, K., and Reichstein, M.: The FLUXCOM ensemble of global land–atmosphere energy fluxes, Sci. Data, 6, 74, https://doi.org/10.1038/s41597-019-0076-8, 2019.

Khan, M. S., Jeon, S. B., and Jeong, M. H.: Gap-filling eddy covariance latent heat flux: Inter-comparison of four machine learning model predictions and uncertainties in forest ecosystem, Remote Sens., 13, 4976, https://doi.org/10.3390/rs13244976, 2021.

Kim, Y., Johnson, M. S., Knox, S. H., Black, T. A., Dalmagro, H. J., Kang, M., Kim, J., and Baldocchi, D.: Gap-filling approaches for eddy covariance methane fluxes: A comparison of three machine learning algorithms and a traditional method with principal component analysis, Global Change Biol., 26, 1499–1518, https://doi.org/10.1111/gcb.1484, 2020.

LeDell, E., and Poirier, S.: H2O AutoML: Scalable Automatic Machine Learning, 7th ICML Workshop on Automated Machine Learning, https://www.semanticscholar.org/paper/H2O-AutoML:-Scalable-Automatic-Machine-Learning-LeDell-Poirier/22cba8f244258e0bba7ff4bb70c4e5b5ac3e2382, 2020.

Lu, J., Wang, G. J., Chen, T. X., Li, S. J., Hagan, D. F. T., Kattel, G., Peng, J., and Jiang, T.: A harmonized global land evaporation dataset from model-based products covering 1980–2017, Earth Syst. Sci. Data, 13, 5879–5898, https://doi.org/10.5194/essd-13-5879-2021, 2021.

Li, Y., Liu, X. N., Zhang, X. G., Gu, X. B., Yu, L. Y., Cai, H. J., and Peng, X. B.: Using solar-induced chlorophyll fluorescence to predict winter wheat actual evapotranspiration through machine learning and deep learning methods, Agric. Water Manag., 309, 109322, https://doi.org/10.1016/j.agwat.2025.109322, 2025a.

Li, S. L., Zhu, P. Y., Song, N., Li, C.. X., and Wang, J. L.: Regional Soil Moisture Estimation Leveraging Multi-Source Data Fusion and Automated Machine Learning, Remote Sensing, 17(5), 837, https://doi.org/10.3390/rs17050837, 2025b.

Li, Z. Q., Yu, G. R., Wen, X. F., Zhang, L. M., Ren, C. Y.: Energy balance closure at ChinaFLUX sites. Science in China Series D: Earth Sciences, 48(Supp. I), 51-62. https://www.sciengine.com/doi/pdf/cc93c4143a194c4cba3ef683920a2791?ipInfo=113.140.84.106, 2005.

Liu, Y. J., Wang, W., Zhao, T. Q., and Huo, Z.Y.: Performance Evaluation and Spatiotemporal Dynamics of Nine Reanalysis and Remote Sensing Evapotranspiration Products in China, Remote Sensing, 17(11), 1881, https://doi.org/10.3390/rs17111881, 2025.

MODIS LAI/FPAR Product Team, MODIS Collection 6.1 (C6.1) LAI/FPAR Product User's Guide. LP DAAC, User Guide. https://lpdaac.usgs.gov/documents/926/MOD15_User_Guide_V61.pdf, 2020.

Myneni, R. B., Hoffman, S., Knyazikhin, Y., Privette, J. L., Glassy, J., Tian, Y., Wang, Y., Song, X., Zhang, Y., Smith, G. R., Lotsch, A., Friedl, M., Morisette, J. T., Votava, P., Nemani, R. R., and Running, S. W.: Global products of vegetation leaf area and fraction absorbed PAR from year one of MODIS data, Remote Sens. Environ., https://doi.org/10.1016/S0034-4257(02)00074-3, 2002.

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

Papale, D.: Ideas and perspectives: enhancing the impact of the FLUXNET network of eddy covariance sites, Biogeosciences, 17(22), 5587-5598, https://doi.org/10.5194/bg-17-5587-2020, 2020.

Pastorello, G., Trotta, C., Canfora, E., et al.: The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data, Scientific Data, 7, 225, https://doi.org/10.1038/s41597-020-0534-3, 2020.

Qi, D. H., Fei, X. H., Song, Q. H., Zhang, Y. P., Sha, L. Q., Liu, Y. T., Zhou, W. J., Lu, Z. Y., Fan, Z. X.: A dataset of carbon and water fluxes observation in subtropical evergreen broad-leaved forest in Ailao Shan from 2009 to 2013. China Scientific Data, 6(1), https://doi.org/10.11922/csdata.2020.0089.zh, 2021.

Qian, L., Zhang, Z. T., Wu, L. F., Fan, S. S., Yu, X. J., Liu, X. G., Ba, Y. L., Ma, H. J., Wang, Y.C.: High uncertainty of evapotranspiration products under extreme climatic conditions. J. Hydrol., 626, 130332, https://doi.org/10.1016/j.jhydrol.2023.130332, 2023.

Qian, L., Yu, X. J., Wu, L. F., Zhang, Z. T., Fan, S. L., Du, R. Q., Liu, X. G., Yang, Q. L., Qiu, R. J., Cui, Y. K., Huang, G. M., and Wang, Y. C.: Improving high uncertainty of evapotranspiration products under extreme climatic conditions based on deep learning and ERA5 reanalysis data, J. Hydrol., 641, 131755, https://doi.org/10.1016/j.jhydrol.2024.131755, 2024a.

Qian, L., Wu, L. F., Zhang, Z. T., Fan, J. L., Yu, X. J., Liu, X. G., Yang, Q. L., Cui, Y. K.: A gap-filling method for daily evapotranspiration of global flux data sets based on deep learning, J. Hydrol., 641, 131787, https://doi.org/10.1016/j.jhydrol.2024.131787, 2024b.

Qian, L., Yu, X. J., Wu, L. F., Cui, Y. K., Zhang, Z. T., Chen, J. Y., Ye, S. M., Bai, X. Q., Liu, X. G., Li, S. E., and Qiu, R. J.: A benchmark dataset for half-hourly evapotranspiration estimation in China from 2000 to 2024 (Version v1). Zenodo [data set], https://doi.org/10.5281/zenodo.18194590, 2026.

Rodell, M., Beaudoing, H. K., L'Ecuyer, T. S., Olson, W. S., Famiglietti, J. S., Houser, P. R., Adler, R., Bosilovich, M. G., Clayson, C. A., Chambers, D., Clark, E., Fetzer, E. J., Gao, X., Gu, G., Hilburn, K., Huffman, G. J., Lettenmaier, D. P., Liu, W. T., Robertson, F. R., Schlosser, C. A., Sheffield, J., and Wood, E. F.: The observed state of the water cycle in the early twenty-first century, J. Clim., 28, 8289–8318, https://doi.org/10.1175/JCLI-D-14-00555.1, 2015.

Running, S. W., Nemani, R. R., Heinsch, F. A., Zhao, M., Reeves, M., Hashimoto, H.: A continuous satellite-derived measure of global terrestrial primary production. BioScience, 54(6), 547-560, https://doi.org/10.1641/0006-3568(2004)054[0547:ACSMOG]2.0.CO;2, 2004.

Tang, R. L., Peng, Z., Liu, M., Li, Z. L., Jiang, Y. Z., Hu, Y. X., Huang, L. X., Wang, Y. Z., Wang, J. R., Jia, L., Zheng, C. L., Zhang, Y. Q., Zhang, K., Yao, Y. J., Chen, X. L., Xiong, Y. J., Zeng, Z. 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.

Trenberth, K. E., Fasullo, J. T., and Kiehl, J.: Earth's global energy budget, Bull. Am. Meteorol. Soc., 90, 311–323, https://doi.org/10.1175/2008BAMS2634.1, 2009.

Wilson, K., Goldstein, A., Falge, E., et al.: Energy balance closure at FLUXNET sites, Agricultural and Forest Meteorology, 113(1-4), 223-243, https://doi.org/10.1016/S0168-1923(02)00109-0, 2002.

Xu, C. Y., and Singh, V. P.: Evaluation of three complementary relationship evapotranspiration models by water balance approach to estimate actual regional evapotranspiration in different climatic regions, J. Hydrol., 308, 105–121, https://doi.org/10.1016/j.jhydrol.2004.10.024, 2005.

Xu, Z. W., Liu, S. M., Li, X., Xu, T. R., Zhu, Z. L.: Water vapor-heat-carbon fluxes and meteorological observation matrix dataset in 2012 over Zhangye oasis-desert area, China Scientific Data, 8(3), https://doi.org/10.11922/11-6035.csd.2023.0108.zh, 2023.

Xu, Z. W., Liu, S. M., Che, T., Ren, Z. G., Tan, J, L., Zhang, Y.: A dataset of carbon and water vapor fluxes and meteorological observations in the middle and lower reaches of the oasis-desert region of the Heihe river basin from 2013 to 2022, China Scientific Data, 9(4). https://doi.org/10.11922/11-6035.csd.2024.0099.zh, 2024.

Yu, L., Qiu, G. Y., Yan, C., Zhao, W., Zou, Z., Ding, J., Qin, L., and Xiong, Y.: A global terrestrial evapotranspiration product based on the three-temperature model with fewer input parameters and no calibration requirement, Earth Syst. Sci. Data, 14, 3673–3691, https://doi.org/10.5194/essd-14-3673-2022, 2022.

Yu, G. R., Wen, X. F., Sun, X. M., Tanner, B. D., Lee, X. H., Chen, J. Y.: Overview of ChinaFLUX and evaluation of its eddy covariance measurement, Agricultural and Forest Meteorology, 137(3-4), 125-137, https://doi.org/10.1016/j.agrformet.2006.02.011, 2006.

Zhao, M. Y., Yang, Y., Weng, G. Y., He, W., Yang, H., Nguyen, N. T., Wang, J. Q., Liu, S., Chen, J. Y., Lei, X. H., Ma, T., Huang, Z. Y., and Xu, P. P.: Fusing Enhanced Flux Measurements and Multi-Source Satellite Observations to Improve GPP Estimation for the Qinghai–Tibet Plateau Based on AutoML Techniques, Remote Sensing, 18(1), 130, https://doi.org/10.3390/rs18010130, 2026.

Zhang, K., Zhu, G., Ma, N., Chen, H., and Shang, S.: Improvement of evapotranspiration simulation in a physically based ecohydrological model for the groundwater–soil–plant–atmosphere continuum, J. Hydrol., 613, 128440, https://doi.org/10.1016/j.jhydrol.2022.128440, 2022.

Zhang, L. L., Vrieling, A., Marshall, M., and Nelson, A.: A macroscale evapotranspiration benchmark based on spatial and temporal enhancements to the Budyko framework, J. Hydrol., 661, 133625, https://doi.org/10.1016/j.jhydrol.2025.133625, 2025.

Zheng, H., Yu, G. R., Zhu, X. J., et al.: A dataset of actual evapotranspiration and water use efficiency of typical terrestrial ecosystems in China (2000–2010), China Scientific Data, 4(1), https://doi.org/10.11922/csdata.2018.0034.zh, 2019.

Zhu, S. Y., 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.

Zou, L., Zhan, C., Xia, J., Wang, T., and Gippel, C. J.: Implementation of evapotranspiration data assimilation with catchment scale distributed hydrological model via an ensemble Kalman filter, J. Hydrol., 549, 685–702, https://doi.org/10.1016/j.jhydrol.2017.04.036, 2017.

Zuo, L. F., Zou, L., Xia, J., Zhang, L. P., Cao, H., and She, D. J.: Multi-scale analysis of six evapotranspiration products across China: Accuracy, uncertainty and spatiotemporal pattern, J. Hydrol., 650, 132516, https://doi.org/10.1016/j.jhydrol.2024.132516, 2025.

Gijsbers, P., LeDell, E., Thomas, J., Poirier, S., Bischl, B., and Vanschoren, J.: An Open Source AutoML Benchmark, arXiv [preprint], https://doi.org/10.48550/arXiv.1907.00909, 2019.

LeDell, E., Poirier, S., et al.: H2O AutoML User Guide, H2O.ai, https://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html, 2021.

Ferreira, J. P., et al.: Machine learning and automated machine learning in environmental modelling: A review and benchmark, Environmental Modelling & Software, https://doi.org/10.1016/j.envsoft.2021.105126, 2021.

Hutter, F., Kotthoff, L., and Vanschoren, J.: Automated Machine Learning: Methods, Systems, Challenges, Springer Nature, Cham, Switzerland, https://doi.org/10.1007/978-3-030-05318-5, 2019.

Bodini, M.: Daily Streamflow Forecasting Using AutoML and Remote-Sensing-Estimated Rainfall Datasets in the Amazon Biomes, Signals. 5(4), 569-589, https://doi.org/10.3390/signals5040037, 2023.

Madni, H.A., Umer, M., Ishaq, N., et al.: Water-Quality Prediction Based on H2O AutoML and Explainable AI Techniques, Water, 15(3), 475, https://doi.org/10.3390/w15030475, 2023.

Mahabbati, A., Beringer, J., Leopold, M., et al.: A comparison of gap-filling algorithms for eddy covariance fluxes and their drivers, Geoscientific Instrumentation, Methods and Data Systems, 10(1), 123-140. https://doi.org/10.5194/gi-10-123-2021, 2021.

Vekuri, H., Tuovinen, J., Kulmala, L., et al.: A widely-used eddy covariance gap-filling method creates systematic bias in carbon balance estimates, Scientific Reports, 13, 1720, https://doi.org/10.1038/s41598-023-28827-2, 2023.

Jia, B., Xie, Z., Zeng, Y., et al.: Diurnal and Seasonal Variations of CO2 Fluxes and Their Climate Controlling Factors for a Subtropical Forest in Ningxiang, Advances in Atmospheric Sciences, 32, 553-564, https://doi.org/10.1007/s00376-014-4069-4, 2015.

Papale, D., Reichstein, M., Aubinet, M., et al.: Towards a standardized processing of Net Ecosystem Exchange measured with eddy covariance technique: algorithms and uncertainty estimation, Biogeosciences, 3, 571–583, https://doi.org/10.5194/bg-3-571-2006, 2006.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., et al.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146(730), 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., and Thépaut, J.-N.: ERA5-Land: A state-of-the-art global reanalysis dataset for land applications, Earth System Science Data, 13, 4349–4383, https://doi.org/10.5194/essd-13-4349-2021, 2021.