Response to reviewer #1 Comments - Round 1

Reviewer general comments: General comments: Thanks to the authors for their valuable contribution to the in-situ flux data collection for Latent heat flux data from various Chinese ecosystems for major climate zones.

The manuscript describes a continuous dataset based on half-hourly ET data from eddy covariance measurements complemented by site-specific ML approaches to fill data gaps and extend the time series beyond the observational periods.

The article is appropriate to support the publication of the data set, but could be shortened in some parts as exemplary indicated below.

The dataset is unique and useful in terms of duration and treatment for Chinese ET flux data. The public availability in open access repositories of such a dataset is relevant even if the creation of the gap-filled and prolonged dataset might be repeated in case observed data together with the software codes were available.

The aggregates for daily, monthly and yearly products could be omitted, as those time series could easily be derived by the users. In case those files are published, I’ d recommend to change the naming of the single files within each archive: each file name should contain the time resolution as is done for the compressed archives.

Within a framework as established for this manuscript, I would expect some more information related to uncertainty of the final flux products that could e.g. be derived from random repetitions of the procedures.

data quality: The 25 year long ET time series is of good quality, also input data for variables that drive ET from MODIS and ERA5 seem reasonable.

Author Respond: We sincerely thank the reviewer for the careful reading of our manuscript and for providing highly valuable and constructive comments. We greatly appreciate the reviewer’s positive evaluation of the dataset and its relevance, as well as the insightful suggestions for improving the manuscript. These comments have been extremely helpful in enhancing the scientific rigor, clarity, and overall readability of the paper. In particular, the reviewer’s remarks on data presentation, manuscript structure, and uncertainty assessment have guided us to substantially refine both the content and organization of the manuscript. We have carefully considered all suggestions and revised the manuscript accordingly, resulting in a clearer and more robust presentation of the dataset and methodology.

We agree with the reviewer that the aggregated products at daily, monthly, and yearly scales could, in principle, be derived by users from the half-hourly dataset. However, these aggregated datasets have already been published and archived on the Zenodo platform. Due to the platform’s policy, once a dataset has been officially released, it cannot be modified or removed. Therefore, we are unfortunately unable to delete these products. We sincerely apologize for this limitation. Nevertheless, we have followed the reviewer’s suggestion to improve clarity by refining the manuscript structure and streamlining the corresponding descriptions to avoid redundancy.

Regarding the uncertainty of the final gap-filled and prolonged flux products, we fully agree with the reviewer on its importance. In response, we have added a dedicated subsection, “4.2 Uncertainty assessment of gap-filling performance,” in the revised manuscript. In this section, we quantify the uncertainty of the AutoML-based gap-filling results using 20 repeated experiments with different random data splits and gap realizations. We also introduce Figure 12, which presents the distributions of CC, RMSE, and PBias under different artificial gap scenarios, thereby providing an empirical estimate of model uncertainty and demonstrating the robustness of the proposed framework.

All other specific comments raised by the reviewer have been addressed in detail in the following point-by-point responses.

We once again thank the reviewer for the insightful and constructive feedback, which has significantly improved the quality of our manuscript.

Specific comments:

Detailed comments 1

What about uncertainty estimates of the gap-filled ET fluxes e.g. based on random repetitions of the procedures?

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we fully agree that uncertainty estimation is essential for evaluating the reliability of the gap-filled latent heat flux (LE) data.

Following the reviewer's recommendation, we have added a dedicated discussion on gap-filling uncertainty in the revised manuscript as Sect. 4.2, "Uncertainty assessment of gap-filling performance." In this new section, the uncertainty of the AutoML-based gap-filling results is quantified using 20 repeated experiments under each artificial gap scenario, with different random data splits and gap realizations. This repeated-experiment design provides an empirical estimate of the uncertainty associated with the gap-filling procedure.

To present these results clearly, we have added Fig. 12, which summarizes the distributions of CC, RMSE, and PBias for the four artificial gap scenarios (30 min, 1 day, 7 day, and 30 day) using boxplots. The results show that:

1. the model uncertainty is relatively low for short gaps, with performance metrics remaining highly concentrated across repetitions;

2. uncertainty increases gradually with gap length, as indicated by broader distributions of CC, RMSE, and PBias;

3. even under the 30 day continuous-gap scenario, the overall spread remains limited, indicating that the framework retains good robustness under repeated realizations.

In particular, the repeated experiments show that CC decreases progressively from about 0.91 for the 30 min scenario to about 0.82 for the 30 day scenario, while RMSE increases from about 30–31 W/m^2 to about 36–39 W/m^2. PBias remains close to zero for short and medium gaps, and although its spread becomes larger under the 30 day scenario, no severe systematic instability is observed.

We believe that this additional analysis substantially strengthens the manuscript by explicitly characterizing the uncertainty of the gap-filling results and by demonstrating the statistical robustness of the proposed framework under random repetitions.

We sincerely appreciate the reviewer's suggestion, which has helped us improve the rigor and completeness of the manuscript.

The new Section 4.2 in ***Discussion*** (**the Line 568 to 591** of the revised manuscript):

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/m2, 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/m2, 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/m2, 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.

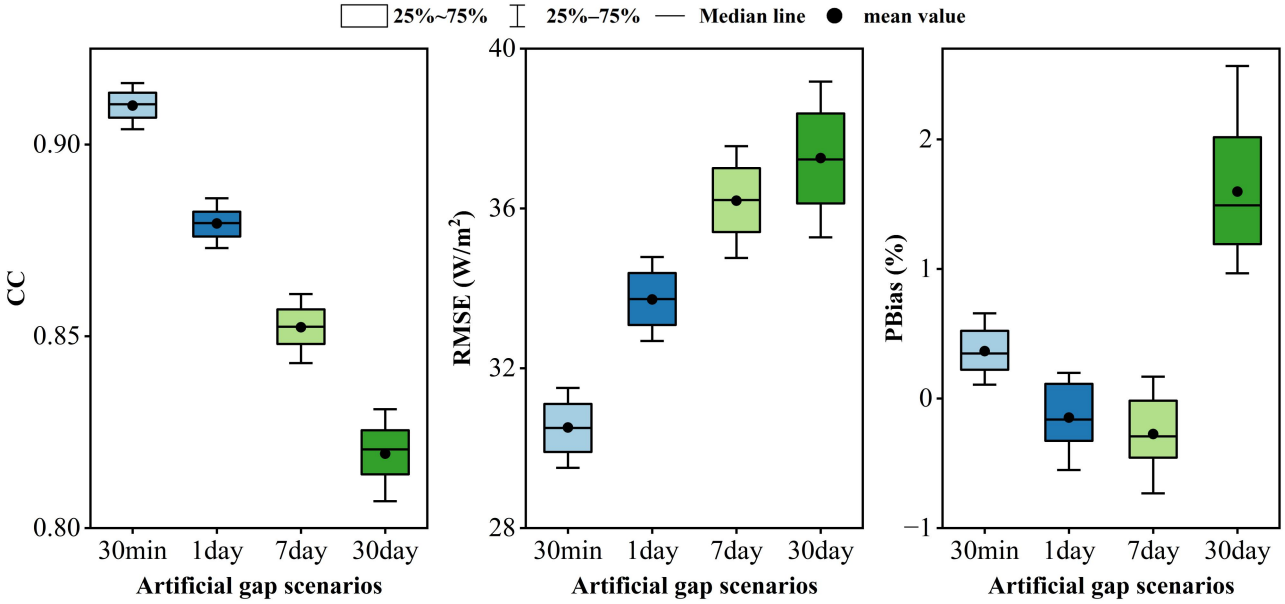

Figure 12: Uncertainty of AutoML-based half-hourly LE gap-filling performance under different artificial gap scenarios (30 min, 1 day, 7 day, and 30 day), estimated from 20 repeated experiments with different random data splits and gap realizations. Boxplots show the distributions of correlation coefficient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

Detailed comments 2

Several references in the text are missing, even though listed in the 'References' section. Please check.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we have carefully reviewed and cross-checked all references in the manuscript. In the revised version, we have thoroughly corrected the reference list to ensure that all references listed in the "References" section are explicitly cited in the main text, and that all in-text citations have corresponding entries in the reference list.

In addition, we have removed unused references and corrected minor inconsistencies to ensure a strict one-to-one correspondence between the text and the reference list.

We appreciate the reviewer's comment, which has helped us improve the overall accuracy and consistency of the manuscript.

Detailed comments 3

Naming of the model: AutoML-H2O or H2O AutoML or only AutoML? Please use the abbreviation consistently.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the naming of the model should be consistent throughout the manuscript. In the revised version, we have standardized the terminology across the entire paper. Specifically, we now use "AutoML" as the unified abbreviation throughout the manuscript for clarity and readability.

To ensure proper definition, the full name is introduced at its first occurrence and in the methodology section as "AutoML of the H2O framework, after which only "AutoML" is used consistently.

This revision improves the consistency and avoids potential confusion for readers.

Specific remarks related to the text:

Detailed comments 4

Line 26: '…conventional methods for long-gap conditions of 7 or 30 days, respectively.'? You might reformulate the sentence as those gap conditions are artificially introduced. Under normal conditions, data gaps vary in duration.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that, in real-world flux observations, data gaps occur with variable durations rather than fixed lengths. In this study, the 7-day and 30-day gap conditions refer to artificial gap scenarios specifically designed to systematically evaluate model performance under different missing-data scales. To avoid potential misunderstanding, we have revised the sentence to explicitly clarify that these gap conditions are introduced in the artificial gap experiments.

This modification improves clarity and better distinguishes between controlled experimental settings and real-world gap characteristics.

The revised version in ***Abstract*** **(the Line 26 to 27** of the revised manuscript, the modified content is displayed in bold):

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.**

Detailed comments 5

Line 30: does the 'strict quality control' relate to the EC data from measurements or to the modelled ET data?

Author Respond: Thank you for pointing out the issues and providing valuable suggestions.

In this study, the term "strict quality control" refers specifically to the eddy covariance (EC) measurements, rather than the modeled ET (LE) data. More precisely, only half-hourly latent heat flux (LE) observations with the highest quality flag (i.e., QA/QC = 0) from the ChinaFlux dataset were selected as reference data for comparison.

To avoid ambiguity, we have revised the sentence accordingly to explicitly indicate that the strict quality control applies to the observed EC data. This clarification ensures a clear distinction between quality-controlled observations and model-generated data.

The revised version in ***Abstract*** (**the Line 30 to 31** of the revised manuscript, the modified content is displayed in bold):

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.

Detailed comments 6

Line 65: 'EC observations provide half-hourly measurements of latent heat flux (LE) in combination with its driving variables,…'. (EC itself delivers only fluxes)

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the original wording was not sufficiently precise. The eddy covariance (EC) system directly provides turbulent flux measurements, whereas the meteorological and environmental driving variables are obtained from concurrent auxiliary measurements at the flux tower sites rather than from the EC system itself.

To address this issue, we have revised the sentence to more accurately distinguish between flux measurements and associated driving variables. This revision improves the technical accuracy of the description.

The revised version in ***Introduction*** (**the Line 66 to 67** of the revised manuscript, the modified content is displayed in bold):

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.

Detailed comments 7

Line 69: 'Taking the FLUXNET2015 dataset as an example,…'

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the expression should be made more complete and precise. Following your suggestion, we have revised the sentence by explicitly referring to the FLUXNET2015 dataset to improve clarity and formal consistency in the manuscript.

The revised version in ***Introduction*** (**the Line 71 to 72** of the revised manuscript, the modified content is displayed in bold):

Taking the FLUXNET2015 dataset as an example, the average missing rate of hourly LE data is approximately 40 %, exceeding 70 % at some sites

Detailed comments 8

Lines 85ff: You claim that Chinese flux sites are underrepresented in integrative flux analysis. Might this also be due to the fact that data from Chinaflux are usually not accessible for non-Chinese researchers? Data sent to the FLUXNET portal should be accessible via the FLUXNET Shuttle (https://data.fluxnet.org/). Also see Papale, D.: Ideas and perspectives: enhancing the impact of the FLUXNET network of eddy covariance sites, Biogeosciences, 17, 5587–5598, https://doi.org/10.5194/bg-17-5587-2020, 2020.

Author Respond: Thank you for this insightful comment and for pointing out the important role of data accessibility in global flux data integration.

We agree that the underrepresentation of Chinese flux sites in global synthesis studies may not only be related to the number and temporal continuity of available sites, but could also be influenced by data availability and accessibility, particularly in the context of international data sharing frameworks such as FLUXNET.

In the original manuscript, our intention was to highlight the limited representation of Chinese sites in existing global analyses. However, we acknowledge that the wording could be interpreted as attributing this solely to scientific or observational limitations. To address this concern, we have revised the text to provide a more balanced perspective by explicitly acknowledging that multiple factors, including data accessibility, may contribute to this issue.

This revision avoids over-attribution and better reflects the complexity of the issue. We appreciate the reviewer's suggestion and have also considered the perspective highlighted in Papale (2020), which emphasizes the importance of open data sharing for enhancing the impact of global flux networks.

The revised version in ***Introduction*** (**the Line 90 to 92** of the revised manuscript, the modified content is displayed in bold):

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 (Liu et al., 2023; Qian et al., 2023). 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.

Detailed comments 9

Line 131 and later line 160: is the quality control related to EC data? What are the site selection criteria?

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the original description of "quality control" and "site selection criteria" was not sufficiently explicit, which may have caused ambiguity.

First, in this study, quality control refers specifically to the eddy covariance (EC) observations. The raw flux data were processed following standard EC data processing procedures, and only high-quality half-hourly latent heat flux (LE) observations were retained for subsequent model training and evaluation.

Second, the site selection criteria were defined to ensure both data reliability and sufficient sample size for model development. As described in Sect. 2.1, the selected sites were required to meet the following conditions:

1. An effective observation period of at least 2 years;
2. At least 10,000 half-hourly LE records available for model training.

To address the reviewer's concern, we have revised the manuscript to explicitly clarify both aspects. Specifically, we now:

Clearly indicate that quality control applies to EC observations;

Provide a brief description of the site selection criteria and refer the reader to Sect. 2.1 for full details.

These revisions improve clarity and ensure better transparency of the data processing and site selection procedures.

The revised version in ***Introduction*** (**the Line 134 to 135** of the revised manuscript, the modified content is displayed in bold):

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.

Detailed comments 10

Lines 158ff: references for the processing and quality assurance steps should be added.

Lines 160 and 161ff: references and details for ChinaFlux and FLUXNET procedures as well as previous studies should be added.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the original manuscript did not provide sufficient references and supporting details for the data processing and quality assurance procedures, particularly with respect to the ChinaFlux and FLUXNET protocols.

In the revised manuscript, we have added a set of relevant references to better support these statements.

First, for the processing and quality assurance procedures of ChinaFlux observations, we added references describing both the general framework and representative site-level implementations, including Yu et al. (2006) and Zheng et al. (2019), as well as several China Scientific Data papers for individual ChinaFlux sites or regional observational matrices (Qi et al., 2021; Xu et al., 2023; Xu et al., 2024). These studies document standard EC data processing and quality control procedures, such as coordinate rotation, WPL correction, outlier screening, and data quality assessment.

Second, for the FLUXNET-related processing framework, we added Pastorello et al. (2020), which provides a comprehensive description of the FLUXNET2015 dataset and the ONEFlux processing pipeline, including standardized quality control, u* threshold estimation, gap-filling, and uncertainty assessment. This reference is particularly appropriate for supporting the statement that our processing standards are compatible with internationally adopted FLUXNET procedures.

Third, to support the statement regarding the reliability of ChinaFlux observations and energy balance closure performance, we added Li et al. (2005) and Wilson et al. (2002). These studies systematically evaluated the energy balance closure characteristics of ChinaFLUX and FLUXNET sites, respectively, and showed that the closure levels of ChinaFlux sites fall within the internationally reported range for eddy covariance observations.

Based on your suggestion, we have revised the text to explicitly include these references and clarify the basis for the methodological statements. The revised sentences now read:

“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 and FLUXNET (Yu et al., 2006; Qi et al., 2021; Xu et al., 2023; Xu et al., 2024; Zheng et al., 2019; 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.”

We appreciate this comment, which helped us improve the methodological transparency and literature support of the manuscript.

The revised version in ***Data and methodology*** (**the Line 134 to 135** of the revised manuscript, the modified content is displayed in bold):

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.

*The **Suplementary References**:*

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.

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.

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.

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.

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.

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.

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.

Detailed comments 11

Line 178: did you use the data from the 10 sites with pre-gap-filled time series in the same way as the data from the other 40 sites? So only artificial gaps introduced? Are the gap-filled data of those sites marked as such? If these gap-filled data are used for training, 'no new information is generated'.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the use of sites with pre-gap-filled time series requires careful clarification, particularly regarding their role in model training and the potential concern that "no new information is generated."

In this study, the 10 sites in question provide continuous LE time series that have already been gap-filled within their observation periods. For these sites, we treat the provided data as complete observations (i.e., with no missing values during the observation period). Importantly, we do not use any externally gap-filled values as training targets in a way that would propagate prior gap-filling assumptions into our model.

To ensure methodological consistency across all sites, we adopt a unified strategy:

1. Artificial gaps are introduced for all sites, including these 10 sites, following the same experimental design used for the other 40 sites. This allows the model to be trained and evaluated under controlled and comparable gap scenarios.
2. Because these sites are treated as having no missing data within the observation period, no "gap-filled" (F) data are generated for them during this period.
3. The model is then used only for temporal prolongation beyond the observation period, and these extended data are explicitly labeled as P (prolonged).
4. In addition, it is important to emphasize that all models are trained independently at the site level, so the inclusion of these 10 sites does not influence the training or results of other sites.
5. To ensure transparency and user flexibility, these 10 sites are explicitly identified in Table A1, allowing users to include or exclude them depending on their specific research needs.

Based on your suggestion, we have revised the manuscript to clearly describe this treatment and avoid potential misunderstanding. We appreciate this comment, which helped us improve the clarity and transparency of our data processing and modeling strategy.

The revised version in ***Data and methodology*** (**the Line 183 to 190** of the revised manuscript, the modified content is displayed in bold):

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 offer continuous LE time series that have already been gap-filled within their observation periods. **For these 10 sites, the provided time series are treated as complete observations (i.e., with no missing data during the observation period), and no additional gap-filled data are used for model training. Instead, artificial gaps are introduced in the same manner as for the other 40 sites to ensure a consistent training strategy across all sites.** Given the overall scarcity of ChinaFlux data, these sites were retained to allow users flexibility in data selection according to specific research objectives. **It should be noted that, for these sites, no gap-filling (F) data exist within the observation period, and only temporally prolonged data outside the observation period are generated and explicitly flagged as P. In addition, these sites are separately identified (Table A1), allowing users to include or exclude them depending on their specific application needs.**

Detailed comments 12

Lines 182-183: 'The half-hourly LE data form the foundation for subsequent gap-filling,…' is that what you want to say? If yes, please re-formulate accordingly.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions.

Yes, our intention was to indicate that the half-hourly LE observations serve as the basis for subsequent gap-filling, temporal prolongation, and dataset construction. We agree that the original wording could be improved for clarity and precision.

This revision improves the clarity and readability of the statement.

The revised version in ***Data and methodology*** (**the Line 193 to 194** of the revised manuscript, the modified content is displayed in bold):

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.**

Detailed comments 13

Line 193: citation for ERA5-data missing in the text (e.g.: Copernicus Climate Change Service (2022): ERA5-Land hourly data from 1950 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). DOI: 10.24381/cds.e2161bac (Accessed on 11-11-2025)

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that a formal citation for the ERA5-Land dataset should be included to properly acknowledge the data source. Following your suggestion, we have added the recommended reference to the Copernicus Climate Change Service dataset description.

This addition ensures proper attribution and improves the completeness of the data description.

The revised version in ***Data and methodology*** (**the Line 208** of the revised manuscript, the modified content is displayed in bold):

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. 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 **(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).

*The **Suplementary References***:

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.

Detailed comments 14

Line 203: I had no glue on the ERA5-Land data, but is Rn really net solar radiation, which would be Albedo? Instead, Rn is commonly used for net radiation including longwave components. Later in the text (line 571, line 577) Rn is used for net radiation.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that, in the conventional definition, Rn typically refers to net radiation, which includes both shortwave and longwave components. In our manuscript, however, the variable denoted as Rn specifically corresponds to the ERA5-Land variable “surface net solar radiation”, which represents the net shortwave radiation at the surface.

We acknowledge that using the symbol Rn for net solar radiation may cause confusion, especially since Rn is more commonly used to represent total net radiation in the literature. To avoid ambiguity, we have revised the text to explicitly clarify this

definition. This clarification ensures consistency with the ERA5-Land dataset while making the distinction from the conventional definition of net radiation clear to readers.

We appreciate the reviewer's comment, which helped improve the precision and clarity of our variable description.

The revised version in ***Data and methodology*** (**the Line 214 to 215** of the revised manuscript, the modified content is displayed in bold):

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).

Detailed comments 15

Line 213: reference for 'official documentation'?

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the phrase "official documentation" should be supported by explicit references. In the revised manuscript, we have added the official MODIS product user guides for the vegetation index and LAI/FPAR products, which document the scale factors used to convert the original MODIS digital values into physical values. Specifically, the MOD13/MYD13 vegetation index guide provides the scaling information for NDVI, and the MOD15/MYD15 LAI/FPAR guide provides the scaling information for LAI.

This revision makes the data processing procedure more transparent and properly documents the source of the scale factors.

The revised version in ***Data and methodology*** (**the Line 224 to 226** of the revised manuscript):

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).**

*The **Suplementary References (official documentation)***:

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.

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.

Detailed comments 16

Lines 210ff: citation for MODIS products.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the MODIS products used in this study should be properly referenced. In the revised manuscript, we have added citations to the official MODIS data products for both vegetation indices and LAI/FPAR.

Specifically:

The NDVI data derived from MOD13Q1 and MYD13Q1 are now supported by the MODIS Vegetation Indices product reference (Didan, 2015);

The LAI data derived from MOD15A2H and MYD15A2H are now supported by the MODIS LAI/FPAR product reference (Myneni et al., 2015).

We note that Terra and Aqua products belong to the same product family and share the same algorithm framework; therefore, a single reference is provided for each product type (NDVI and LAI), which is consistent with common practice.

This revision ensures proper attribution and improves the completeness of the data description.

The revised version in ***Data and methodology*** (**the Line 223** of the revised manuscript, the modified content is displayed in bold):

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)**.

*The **Suplementary References***:

Didan, K.: MOD13Q1 MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V006 [Data set]. NASA Land Processes Distributed Active Archive Center, https://doi.org/10.5067/MODIS/MOD13Q1.006, 2015.

Myneni, R., Knyazikhin, Y., Park, T.: MOD15A2H MODIS/Terra Leaf Area Index/FPAR 8-Day L4 Global 500m SIN Grid V006 [Data set]. NASA Land Processes Distributed Active Archive Center, https://doi.org/10.5067/MODIS/MOD15A2H.006, 2015.

Detailed comments 17

Line 216: what is the temporal resolution of MODIS products ('relatively low')? And is the assumption of constant NDVI and LAI 'within each compositing period' sufficient for fast growing plants, disturbances etc.?.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions.

First, we agree that the phrase "relatively low temporal resolution" should be specified more clearly. In the revised manuscript, we now explicitly state that the MODIS NDVI (MOD13Q1/MYD13Q1) products are provided as 16-day composites, while the LAI (MOD15A2H/MYD15A2H) products are 8-day composites.

Second, regarding the assumption that NDVI and LAI remain constant within each compositing period, we acknowledge that this is a simplification. This approach is widely adopted in studies integrating remote sensing products with high-frequency flux data, because MODIS compositing procedures are designed to reduce short-term noise (e.g., cloud contamination and atmospheric variability) and to represent average vegetation conditions over the compositing interval (Didan et al., 2015; Myneni et al., 2002; Running et al., 2004).

Recent studies continue to use MODIS LAI/VI as 8-day or 16-day composite inputs, while the official MODIS product documentation makes clear that these variables are designed to represent vegetation conditions over the compositing interval rather than an instantaneous observation. Therefore, assigning one composite value to all sub-daily steps within that interval is a practical temporal alignment assumption, although it may miss rapid vegetation changes.

However, we agree with the reviewer that this assumption may not fully capture rapid vegetation changes, such as those associated with fast-growing crops, disturbances, or management practices. To address this concern, we have added a clarification in the methodology section and explicitly acknowledged this limitation in the discussion section.

These revisions improve both the transparency of the methodological assumption and the discussion of its potential implications.

The revised version in ***Data and methodology*** (**the Line 230 to 236** of the revised manuscript, the modified content is displayed in bold):

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; Running et al., 2004).**

The revised version in ***Discussion*** (**the Line *** to ***** of the revised manuscript, the modified content is displayed in bold):

Nevertheless, several limitations remain. First, model performance still varies across land cover types and climate zones, and uncertainties persist in representing extreme high LE values, particularly in sparsely vegetated or water-limited environments. Second, SHAP analyses indicate a strong model dependence on energy-related variables and vegetation indices. Although bias correction was applied to reanalysis and remote sensing inputs, uncertainties inherent in these drivers may still propagate into the final LE estimates. **In particular, NDVI and LAI derived from MODIS products were assumed to remain constant within each compositing period (8–16 days), which may not fully capture rapid vegetation dynamics (e.g., fast growth, disturbance, or management practices), potentially introducing additional uncertainty at sub-daily scales.** In addition, the temporal prolongation approach assumes that historical statistical relationships remain relatively stable over time, an assumption that warrants further examination under ongoing climate change.

*The **Suplementary References***:

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.

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.

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.

Detailed comments 18

Line 228: re-formulate: 'Each flux tower site was treated…' – it is not the tower that is of relevance.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions,

We agree that the original wording was not sufficiently precise, as the modeling is performed at the site level, rather than on the physical tower itself. To improve clarity and accuracy, we have revised the sentence by replacing "flux tower" with "site".

This modification improves the precision of the terminology and better reflects the actual modeling framework.

The revised version in ***Data and methodology*** (**the Line 246** of the revised manuscript, the modified content is displayed in bold):

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.

Detailed comments 19

Line 230: How was quality screening performed? According to standard FLUXNET and ChinaFlux procedures? See above.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that a clear description of the quality screening procedure is essential for ensuring the transparency and reproducibility of the dataset.

In the revised manuscript, we have explicitly clarified that the quality screening of latent heat flux (LE) observations follows the standard quality control protocols adopted by both FLUXNET and ChinaFlux. Specifically, we retained only high-quality observations with QC/QA flags equal to 0, which correspond to data that have passed all standard processing steps, including coordinate rotation, frequency response correction, WPL correction, friction velocity (u*) filtering, and energy balance closure assessment.

In addition, observations affected by known sources of uncertainty—such as instrument malfunction, low turbulence conditions (e.g., insufficient friction velocity), and poor energy balance closure—were excluded to ensure the reliability of the training dataset.

This clarification ensures that the dataset is based on rigorously screened flux observations and is fully consistent with widely accepted international practices.

The revised version in ***Data and methodology*** (**the Line 248 to 251** of the revised manuscript, the modified content is displayed in bold):

During model development and evaluation, the original LE observations were first subjected to quality screening, following the standard quality control procedures of FLUXNET and ChinaFlux. 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.

Detailed comments 20

Lines 240ff: was the modelling tool adopted from somewhere else (reference!) and what is the contribution of the authors?

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that it is important to clearly acknowledge the origin of the modeling tool and to distinguish our methodological contribution.

In the revised manuscript, we have now explicitly added references to the H2O AutoML framework, including both the original methodological publication (LeDell and Poirier, 2020) and the official documentation (H2O.ai, 2023), in Section 2.4.1. This ensures that the source of the AutoML framework is properly cited and traceable.

We would also like to clarify the contribution of this study. The AutoML framework itself was not developed by the authors; rather, our contribution lies in:

1. Designing a site-specific modeling framework tailored to flux tower latent heat flux (LE) data;

2. Integrating multi-source drivers (ERA5-Land and MODIS) into the AutoML workflow to capture nonlinear controls of evapotranspiration;

3. Extending the application from conventional gap-filling to both gap-filling and temporal prolongation, enabling the reconstruction of long-term continuous half-hourly LE time series;

4. Constructing a national-scale benchmark dataset (2000–2024) based on this framework and systematically evaluating its performance across gap lengths, temporal scales, surface types, and climate zones.

In addition, as noted by the reviewer, we had already acknowledged the H2O AutoML framework and provided its access link in the Acknowledgements section, and expressed our appreciation to the H2O.ai team. However, we apologize for the oversight in not including formal citations in the main text. This has now been corrected in the revised manuscript.

We believe this revision improves both the transparency and academic rigor of the manuscript.

The revised version in ***Data and methodology*** (**the Line 261 to 266** of the revised manuscript, the modified content is displayed in bold):

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**, which has been widely used for regression and prediction tasks in environmental and geoscientific applications (Guo et al., 2024; Li et al., 2025; 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**).

*The **Suplementary References (official documentation)***:

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.

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

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.

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, 2025.

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.

Detailed comments 21

Line 249: Which models were selected by AutoML? Very different models? How many different models?

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that a clearer description of the internal model structure of the AutoML framework is necessary to improve transparency and reproducibility.

The AutoML approach used in this study is based on the H2O AutoML framework, which evaluates a diverse set of machine learning algorithms rather than relying on a single model type. Specifically, the candidate models include:

Gradient Boosting Machine (GBM)

Distributed Random Forest (DRF)

Extreme Gradient Boosting (XGBoost)

Generalized Linear Models (GLM)

Deep Learning Neural Networks (DNN)

Stacked Ensemble models that combine multiple base learners

These models differ substantially in their structures and learning mechanisms, ranging from linear models (GLM) to tree-based ensemble methods (GBM, DRF, XGBoost) and nonlinear neural networks (DNN). This diversity allows the AutoML framework to effectively capture complex nonlinear relationships between latent heat flux (LE) and its driving variables under varying environmental conditions.

Regarding the number of models, the H2O AutoML framework automatically explores multiple model configurations and hyperparameter combinations using a random grid search strategy. In our implementation (**see the provided R code, https://doi.org/10.5281/zenodo.18194590**), no strict upper limit was imposed on the number of models (i.e., max_models = 0), allowing the algorithm to adaptively determine the number of candidate models based on the data characteristics and computational process. Typically, this results in dozens of candidate models being trained and evaluated for each site, although the exact number may vary depending on data availability and model convergence behavior.

Finally, the optimal model for each flux tower site is selected based on validation performance (combined with five-fold cross-validation), ensuring that the final model is both robust and site-specific.

We have revised the manuscript accordingly to explicitly describe the types of models included in AutoML and clarify how the number of models is determined.

The revised version in ***Data and methodology*** (**the Line 274 to 282** of the revised manuscript, the modified content is displayed in bold):

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. **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. 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. 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.

Detailed comments 22

Lines 280ff: ERA5 data are used instead of onsite measured meteorological variables. How do those compare with locally measured data? What about the additional uncertainty?

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we fully agree that the use of reanalysis data instead of in situ meteorological measurements requires careful justification and discussion of the associated uncertainties.

First, regarding the comparison between ERA5-Land data and locally measured meteorological variables, we acknowledge that a comprehensive evaluation across all sites was not feasible. This is primarily because many ChinaFlux sites do not provide complete or continuous records of key meteorological variables (e.g., radiation, temperature, and vapor pressure deficit), which makes a systematic comparison difficult.

However, for a subset of sites where both ERA5-Land data and in situ observations are available, we conducted a preliminary comparison and found generally good agreement between the two datasets. Specifically, for key variables such as air temperature and radiation, the coefficient of determination (R^2) typically ranges from approximately 0.70 to 0.90, indicating that ERA5-Land can reasonably capture the temporal variability of site-scale meteorological conditions.

Second, we acknowledge that the use of ERA5-Land data may introduce additional uncertainty compared to direct in situ measurements, particularly in regions with complex terrain or heterogeneous land surface conditions. To address this concern, we have added a more explicit discussion of this limitation in the revised manuscript. Despite this limitation, we chose to use ERA5-Land data for the following reasons:

1. **Data availability:** Many flux tower sites lack complete meteorological observations, making it impractical to use in situ data consistently across all sites;

2. **Consistency:** ERA5-Land provides spatially and temporally continuous meteorological variables, ensuring a unified input dataset for all sites;

3. **Comparability:** Using the same set of input variables for all models ensures fair comparisons among different gap-filling methods;

4. **Established practice:** The use of reanalysis data in flux gap-filling studies has been widely adopted and validated in previous research.

Furthermore, the machine learning framework used in this study learns the statistical relationships between LE and multiple environmental drivers, which helps mitigate part of the uncertainty introduced by individual input variables.

We have revised both the Methods and Discussion sections to clarify these points and explicitly acknowledge the associated uncertainties.

The revised version in ***Data and methodology*** (**the Line 316 to 325** of the revised manuscript, the modified content is displayed in bold):

To facilitate comparative evaluation of different gap-filling approaches, the traditional marginal distribution sampling (MDS) method and two commonly used machine learning algorithms, random forest (RF) and extreme gradient boosting (XGBoost), were implemented as benchmark methods. 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, 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.

The revised version in ***Discussion*** (**the Line *** to ***** of the revised manuscript, the modified content is displayed in bold):

Nevertheless, several limitations remain. First, model performance still varies across land cover types and climate zones, and uncertainties persist in representing extreme high LE values, particularly in sparsely vegetated or water-limited environments. Second, SHAP analyses indicate a strong model dependence on energy-related variables and vegetation indices. Although bias correction was applied to reanalysis and remote sensing inputs, uncertainties inherent in these drivers may still propagate into the final LE estimates. In particular, NDVI and LAI derived from MODIS products were assumed to remain constant within each compositing period (8–16 days), which may not fully capture rapid vegetation dynamics (e.g., fast growth, disturbance, or management practices), potentially introducing additional uncertainty at sub-daily scales. **And in MDS method, the use of ERA5-Land reanalysis data instead of in situ meteorological measurements may introduce additional uncertainty due to potential biases in representing local microclimatic conditions, especially in complex terrain or heterogeneous landscapes. While ERA5-Land generally shows good agreement with site observations at locations where such data are available, discrepancies at sub-daily scales may still affect the accuracy of LE estimation.** In addition, the temporal prolongation approach assumes that historical statistical relationships remain relatively stable over time, an assumption that warrants further examination under ongoing climate change.

Detailed comments 23

Line 316: the repetitive sentence can be removed here (These prolonged datasets provide the basis for subsequent construction and analysis of multi-temporal-scale LE products.)

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the sentence was redundant. The repetitive sentence has been removed in the revised manuscript accordingly to improve clarity and conciseness.

Detailed comments 24

Lines 332-333: does that mean, in case there would be only one half-hour value missing within a 1 d aggregate, the 1 d value would get the flag F? Hardly any ET time series from EC measurements is complete due to unfavourable atmospheric conditions. As a result, any daily (7 day, monthly, respectively) aggregate is based on a mixture of measured and gap-filled data. How are the aggregates flagged then?

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the quality flagging scheme for aggregated data needs to be clarified.

At the half-hourly scale, the quality flag (QC_LE) is defined as follows:

T: original flux tower observation

F: gap-filled value within the observation period

P: temporally prolonged value outside the observation period

However, for aggregated datasets (daily, monthly, and annual), we do not assign a single categorical quality flag (such as T/F/P) to each aggregated value. Instead, we provide a quantitative indicator of data availability, namely N_obs, which represents the number of half-hourly observations (i.e., original measurements with QC_LE = T) contributing to the aggregated value.

For example, as mentioned in the reviewer's scenario, if only one half-hourly value is missing within a day (i.e., 48 time steps per day), then N_obs = 47. If all half-hourly observations are missing for a given day, then N_obs = 0. The same definition is consistently applied to monthly and annual aggregates.

This approach allows each aggregated value to explicitly reflect the proportion of observed versus reconstructed data, rather than assigning a single categorical flag to a mixed dataset. It also provides users with greater flexibility to apply their own filtering criteria based on data completeness.

We have revised the manuscript accordingly to clarify this point in both the Methods and Data Availability sections.

The revised version in ***Data and methodology*** **(the Line 316 to 325** of the revised manuscript, the modified content is displayed in bold):

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.**

The revised version in ***Data availability*** **(the Line *** to ***** of the revised manuscript, the modified content is displayed in bold):

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.

Detailed comments 25
Line 432: suggestion: '…the 6-year training scenario represents more typical conditions and weather regimes at most sites.'

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the original wording could be improved to better reflect the representativeness of the training data.

Following the reviewer's recommendation, we have revised the sentence to explicitly emphasize that the 6-year training scenario captures more typical environmental conditions and weather regimes at most sites. This modification improves the clarity and scientific accuracy of the description.

The revised version in ***Result*** **(the Line 475** of the revised manuscript, the modified content is displayed in bold):

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.**

Detailed comments 26
Chapters 3.3.2 and 3.3.3: this analysis and the accompanying figures seem to be redundant with very little additional information for a data paper, as the monthly values are just aggregated from the daily values which contain gap-filled data. In addition, figure 11 shows again some typical sites but with very different time periods, varying from only 2 years to 10 years. Why are these examples chosen so different if they are compared to check for seasonal variation?

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the original presentation of Sections 3.3.2 and 3.3.3 contained some redundancy and could be improved for better clarity and conciseness.

Following the reviewer's suggestion, we have made the following revisions:

1. Reduction of redundancy:

The analysis of the daily and monthly time frames has been condensed and combined with the half-hour time frame into a single chapter. This revision removes repetitive descriptions and focuses on the key finding that the prolonged LE data consistently reproduce seasonal patterns and interannual variability across temporal scales.

2. Relocation of figures:

The original Figures 10 and 11 have been moved to the Appendix (now Figs. A1 and A2). This allows us to retain the detailed site-level comparisons for interested readers while keeping the main text more concise and focused on the core results.

3. Clarification of site selection strategy:

We acknowledge that the selected example sites cover different time periods (ranging from approximately 2 to 10 years), which may appear inconsistent at first glance. However, this selection was intentional. The purpose was not only to illustrate seasonal

variation, but more importantly to demonstrate the robustness and generalization ability of the proposed method under diverse conditions. Specifically:

Sites were selected to cover different temporal spans, including both short-term (~2 years) and long-term (~10 years) observation records;

Sites span multiple climate zones and underlying surface types;

In some cases, the availability of flux observations is inherently limited, especially for certain land cover types, which constrains the selection of long-term records.

Therefore, the diversity in time periods reflects realistic data availability conditions and helps evaluate model performance across heterogeneous scenarios, rather than limiting the analysis to uniform-length records.

Additional clarification added to the manuscript:

We have explicitly stated this rationale in the revised manuscript to avoid potential misunderstanding and to better justify the representativeness of the selected examples.

Overall, these revisions improve the clarity, conciseness, and scientific rigor of the manuscript, while preserving the key information needed to evaluate the model performance across multiple temporal scales.

The revised version in ***Result*** (**the Line 510 to 524** of the revised manuscript, the modified content is displayed in bold):

3.3 Demonstration of different scale prolonged time series

To illustrate the effects of temporal prolongation at the half-hourly scale, Fig. 9 presents comparisons between observed half-hourly latent heat flux (LE) and AutoML-based prolonged results at several representative sites spanning different underlying surface types and climate zones. Overall, the prolonged time series reasonably reproduce the diurnal cycles and amplitude characteristics of the observed LE, exhibiting stable temporal continuity and physically consistent structures. At cropland and grassland sites, the prolonged results closely match the observations in terms of the timing of peak values, diurnal variation patterns, and overall magnitude, indicating that the model effectively captures radiation-driven diurnal processes. For forest and wetland sites, where LE amplitudes are larger and short-term fluctuations are more complex, the prolonged series still track the main observed variations well, with only minor deviations occurring at a limited number of extreme peaks. In contrast, at sparsely vegetated sites such as desert and shrubland, LE magnitudes are relatively small and temporal variability is more irregular. Under these conditions, the prolonged results tend to be more conservative in representing isolated extreme high values but remain consistent with observations in terms of overall trends and diurnal rhythms. Similar patterns are observed across different climate zones, particularly in arid and high-altitude regions, where uncertainty is generally higher. Nevertheless, no evident nonphysical jumps or structural distortions are found in the prolonged time series. Overall, the half-hourly time series examples in Fig. 9 demonstrate that the proposed prolongation framework is able to stably reproduce high-frequency LE variability and diurnal cycle structures at the half-hourly scale, thereby providing a reliable basis for subsequent aggregation to daily and monthly timescales.

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 that the model reliably captures evapotranspiration processes controlled by radiation conditions and vegetation seasonality. 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, suggesting that the model maintains robust temporal stability when transitioning across scales.

Detailed comments 27

Line 515: what exactly is 'official ChinaFlux observations'? Please add reference!

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the term "official ChinaFlux observations" was ambiguous and could lead to misunderstanding.

1. **Clarification of terminology:** In the revised manuscript, we have replaced the term "official ChinaFlux observations" with"the original eddy covariance observations from the 50 ChinaFlux sites used in this study (see Sect. 2.1)".This explicitly refers to the same dataset described in Sect. 2.1, where the data sources, site selection criteria, and data processing procedures are clearly documented. This modification ensures consistency in terminology throughout the manuscript and avoids any ambiguity regarding the reference dataset.

2. **On the purpose and necessity of Section 4.1:** We would also like to clarify the role of this comparison within the overall structure of the manuscript. The preceding sections (Sect. 3) primarily evaluate model performance using: artificial gap scenarios, cross-validation strategies, and temporal extrapolation consistency tests. While these evaluations demonstrate the internal robustness and stability of the gap-filling and prolongation framework, they do not fully address a key question for a data paper: whether the final reconstructed dataset preserves the statistical characteristics of the original flux observations.

3. Section 4.1 is therefore designed as an independent, observation-based consistency assessment, with the following specific purposes:

To verify that the reconstructed dataset does not introduce systematic bias relative to the original flux tower observations;

To evaluate consistency across multiple temporal scales (half-hourly to annual), which is essential for long-term applications;

To assess robustness across different land cover types and climate zones, given the strong heterogeneity of evapotranspiration processes in China;

To demonstrate that the dataset retains the physical and statistical properties of EC measurements, even after gap-filling and temporal prolongation.

This step is particularly critical because flux observations are known to contain substantial data gaps and uneven temporal coverage. Without such a comparison, it would be difficult to ensure that the reconstructed dataset remains faithful to the original observations rather than introducing artifacts.

4. **Additional clarification added in the manuscript:** To improve clarity, we have revised the wording in Sect. 4.1 to explicitly state the data source and its connection to Sect. 2.1, ensuring that readers can clearly trace the origin of the reference observations.

The revised version in ***Discussion*** **(the Line 530 to 531** of the revised manuscript, the modified content is displayed in bold):

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).**

Detailed comments 28

Line 547: what is meant with 'stratification' in the context?

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the term "stratification" was not sufficiently clear in this context and could potentially lead to misunderstanding.

In the revised manuscript, we have replaced the term "stratifications" with "grouped analyses based on underlying surface types (Fig. 12) and climate zones (Fig. 13)". This modification explicitly clarifies that the SHAP analyses were conducted by grouping the full sample dataset according to:

different underlying surface types (e.g., cropland, forest, grassland, etc.), and

different climate zones across China.

No statistical stratification procedure (e.g., formal sampling stratification) was applied; rather, the grouping was performed to facilitate interpretation of model behavior under different environmental conditions.

The corresponding sentence has been revised to improve clarity and explicitly indicate the grouping criteria, thereby avoiding ambiguity in terminology.

We appreciate the reviewer's suggestion, which has helped us improve the clarity and precision of the manuscript.

The revised version in ***Discussion*** **(the Line 563 to 564** of the revised manuscript, the modified content is displayed in bold):

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 12 and 13 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.

Detailed comments 29

Lines 560ff: Especially for desert and shrubland, evaporation becomes more dominant compared to transpiration. So it is clear that vegetation-related variables explain less variability. This point might be considered here as well.

Author Respond: Thank you for this insightful and physically meaningful comment. We fully agree with the reviewer's interpretation.

In arid and semi-arid ecosystems such as desert and shrubland, evapotranspiration is indeed dominated by evaporation rather than transpiration, due to sparse vegetation cover and limited plant activity. As a result, vegetation-related variables (e.g., NDVI

and LAI) contribute less to explaining the variability of latent heat flux (LE), while energy input and atmospheric demand become the primary controlling factors.

Following the reviewer's suggestion, we have incorporated this important physical interpretation into the revised manuscript. Specifically, we have added a sentence explicitly linking the reduced importance of vegetation-related variables to the dominance of evaporation processes in these ecosystems.

This addition strengthens the physical interpretability of the SHAP analysis results and better aligns the data-driven findings with established ecohydrological understanding.

We appreciate the reviewer's comment, which has helped improve the clarity and scientific rigor of the discussion.

The revised version in ***Discussion*** **(the Line 580 to 582** of the revised manuscript, the modified content is displayed in bold):

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.

Detailed comments 30

Lines 577ff: same as above, more soil is exposed, so more evaporation compared to dense canopy covers.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we fully agree with the reviewer's interpretation. In arid and semiarid climate zones, vegetation cover is generally sparse, which results in a larger proportion of exposed soil surface. Under such conditions, soil evaporation becomes more dominant relative to plant transpiration, in contrast to densely vegetated ecosystems where transpiration plays a larger role. Consequently, vegetation-related variables (e.g., NDVI and LAI) have reduced explanatory power for latent heat flux (LE), while energy-related and atmospheric variables become more influential.

Following the reviewer's suggestion, we have incorporated this explanation into the revised manuscript to explicitly link the observed decrease in vegetation-related importance to the dominance of soil evaporation processes in sparsely vegetated regions. This revision further strengthens the physical interpretation of the SHAP results and improves the consistency between the data-driven findings and established ecohydrological understanding.

We appreciate the reviewer's valuable comment, which has helped enhance the clarity and scientific rigor of the discussion.

The revised version in ***Discussion*** **(the Line 597 to 599** of the revised manuscript, the modified content is displayed in bold):

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.

Figures: Readers should be able to interpret figures with the figure description. Most figures need more descriptive text:

Detailed comments 31
Fig. 1b) more description needed, **Fig. 1d) length of observation periods (in years) for all sites**

Author Respond: Thank you for pointing out the issues and providing valuable suggestions

Regarding Fig. 1b: We agree that the original description of Fig. 1b was not sufficiently detailed. In the revised manuscript, we have clarified both the variables shown and the meaning of the dual axes.

Specifically, Fig. 1b presents the distribution of site-level missing data proportions of half-hourly LE observations. The figure contains two complementary components:

The histogram (left axis) represents the number of sites within different intervals of missing data percentage, providing an overview of how data gaps are distributed across the network.

The line plot (right axis) shows the missing data percentage for each individual site, indexed by Site_ID (as listed in Table A1), allowing the identification of variability among sites.

This combined representation enables both a statistical summary and a site-specific view of data completeness, which is important for understanding the heterogeneity of data gaps prior to gap-filling.

Regarding Fig. 1d: Following the reviewer's suggestion, we have explicitly clarified that Fig. 1d shows the distribution of observation period lengths in years for all sites, thereby avoiding ambiguity in units.

Improvement to manuscript clarity: These revisions provide a clearer and more precise description of Fig. 1, particularly for Fig. 1b, ensuring that readers can correctly interpret both the statistical distribution and site-level characteristics of missing data.

We appreciate the reviewer's helpful comments, which have improved the clarity and completeness of the manuscript.

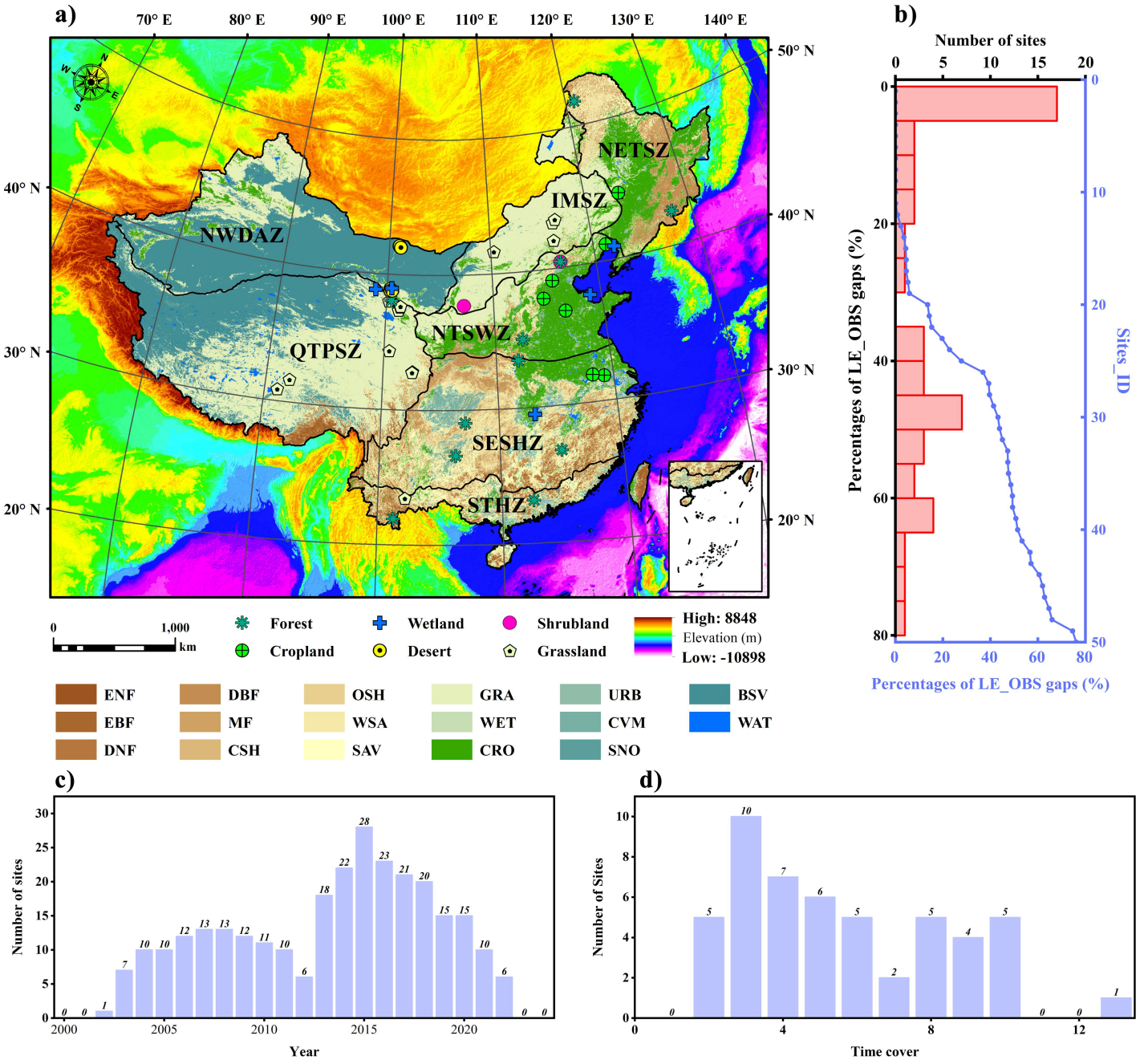

Figure 1: Spatial distribution of the selected ChinaFlux sites (a) and data coverage characteristics of half-hourly LE observations (b–d). Panel (b) shows the distribution of LE data gap percentages across sites, panel (c) presents the number of available sites by year, and panel (d) summarizes the length of observation periods for all sites.

The revised version in ***Data and methodology*** **(the Line 191 to 197** of the revised manuscript, the modified content is displayed in bold):

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.

Detailed comments 32

Figure 2: a bit overwhelming, but a good overview still. I don't see QA/QC for EC-data (which contributes to gaps). The figures for comparison with MDS and ML methods and also the ones for performance do not add any value due to their small size. The text is not readable and legends are missing. Even if the content becomes clear for the results in the lower right corner, instead of T, F, and P for true, filled and prolonged you might use additional colour indication as T, F and P are not in the figures anyway.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we appreciate the reviewer's positive evaluation of Fig. 2 as an overview and fully agree that its clarity and readability can be further improved.

In the revised manuscript, we have carefully modified Fig. 2 and its presentation, with the following improvements:

1. **Clarification of QA/QC procedures:** We agree that the quality control of EC data is a critical component, particularly as it directly contributes to data gaps. In the revised figure and accompanying description, we have explicitly indicated that only quality-controlled observations (QC/QA = 0) are retained for subsequent analysis, thereby clarifying the role of QA/QC in the data processing workflow.
2. **Removal of low-resolution subpanels:** As suggested, the subpanels illustrating comparisons with MDS and machine learning (ML) methods, as well as those related to model performance, were too small to be informative. These panels have been removed from Fig. 2, and their content is now briefly described in the main text, where they can be presented more clearly and without visual limitations.
3. **Improvement of visual clarity and legend design:** To enhance readability, we have simplified the figure layout and improved the graphical presentation. In particular, the previous textual labels ("T", "F", and "P") have been replaced by distinct color-coded legend elements.

This change improves interpretability and avoids reliance on abbreviations that were not clearly visible in the original figure.

Overall, these revisions significantly improve the clarity, readability, and informational value of Fig. 2, while preserving its role as a concise overview of the data processing framework.

We thank the reviewer for these helpful suggestions, which have led to a clearer and more effective presentation.

The revised ***Figure 2***:

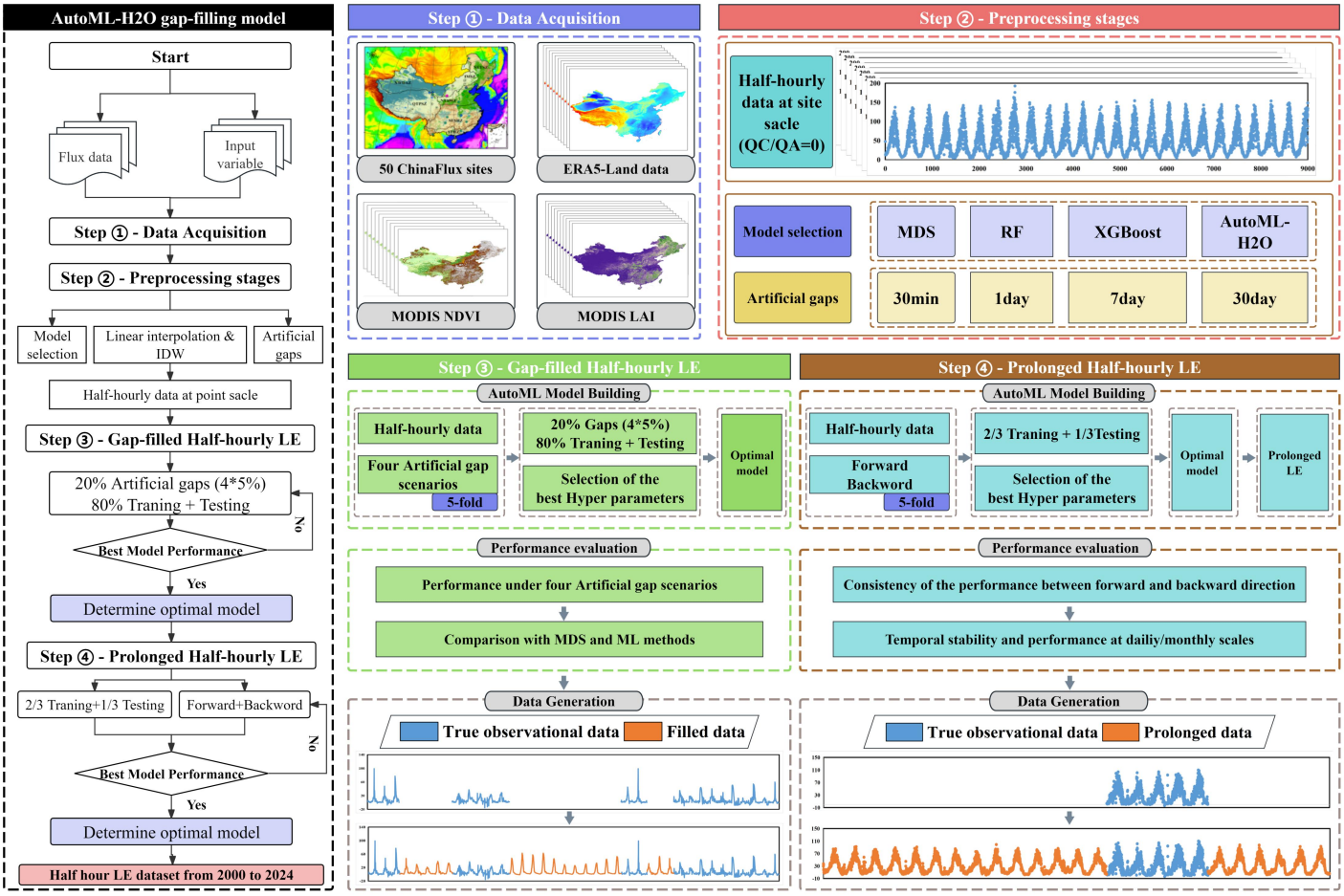

Figure 2: Flowchart and schematic of the gap-filling and prolongation framework for LE_OBS data.

Detailed comments 33

Fig. 5 and 6: what is the measure for the significant bias marked by the blue boxes?

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we would like to clarify that the blue dashed boxes shown in Figures 5 and 6 do not correspond to a predefined quantitative metric or formal statistical test of significance. Instead, they are intended as visual annotations to highlight time periods where noticeable deviations between the MDS gap-filling results and the reference observations (LE_OBS) can be clearly identified in the time series.

These highlighted segments were selected based on the following considerations:

persistent or systematic divergence from observed values,

apparent overestimation or underestimation patterns, and

deviations that are visually distinct compared to surrounding periods.

The purpose of these annotations is to guide the reader's attention to representative cases where the differences between methods are most evident, thereby improving the interpretability of the figure. They are not meant to imply statistical significance in a strict sense.

To avoid potential misunderstanding, we have revised the figure captions to explicitly state that the blue dashed boxes represent qualitative visual highlights rather than results based on a formal quantitative criterion.

We appreciate the reviewer's comment, which has helped us improve the clarity and rigor of the figure presentation.

The revised ***title*** of ***Figure 5*** and ***Figure 6*** (the modified content is displayed in bold):

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.**

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.**

Detailed comments 34

Fig. 7, a) and b): y-axis needs legend. More explanatory test in the figure description is needed, e.g. for 'Relative density'

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that additional clarification and improved labeling are necessary to enhance the readability and interpretability of Fig. 7.

In the revised manuscript, we have made the following improvements:

1. **Addition of y-axis label:** We have added a consistent y-axis label, "Prolonged LE (W/m^2)", to both Fig. 7a and Fig. 7b. This ensures that the plotted variable is clearly defined and consistent with the x-axis ("Observed LE"), improving the clarity of the comparison between observed and prolonged latent heat flux.

2. **Clarification of "Relative density":** We agree that the term "relative density" required further explanation. In the revised figure caption, we have added a clear description indicating that the color scale represents the normalized point density of the scatter distribution, where warmer colors (e.g., red/yellow) indicate regions with a higher concentration of data points, and cooler colors (e.g., blue) indicate lower density. This helps readers better interpret the distribution patterns and the agreement between observed and prolonged LE.

3. **Improved figure caption:** The caption of Fig. 7 has been revised to include the above clarification and to provide a more complete explanation of panels (a)–(f), ensuring that all elements of the figure are clearly described.

These revisions improve both the visual clarity and the interpretability of Fig. 7.

We appreciate the reviewer's suggestion, which has helped us enhance the overall quality of the figure presentation.

The revised ***Figure 7*** and ***its title***:

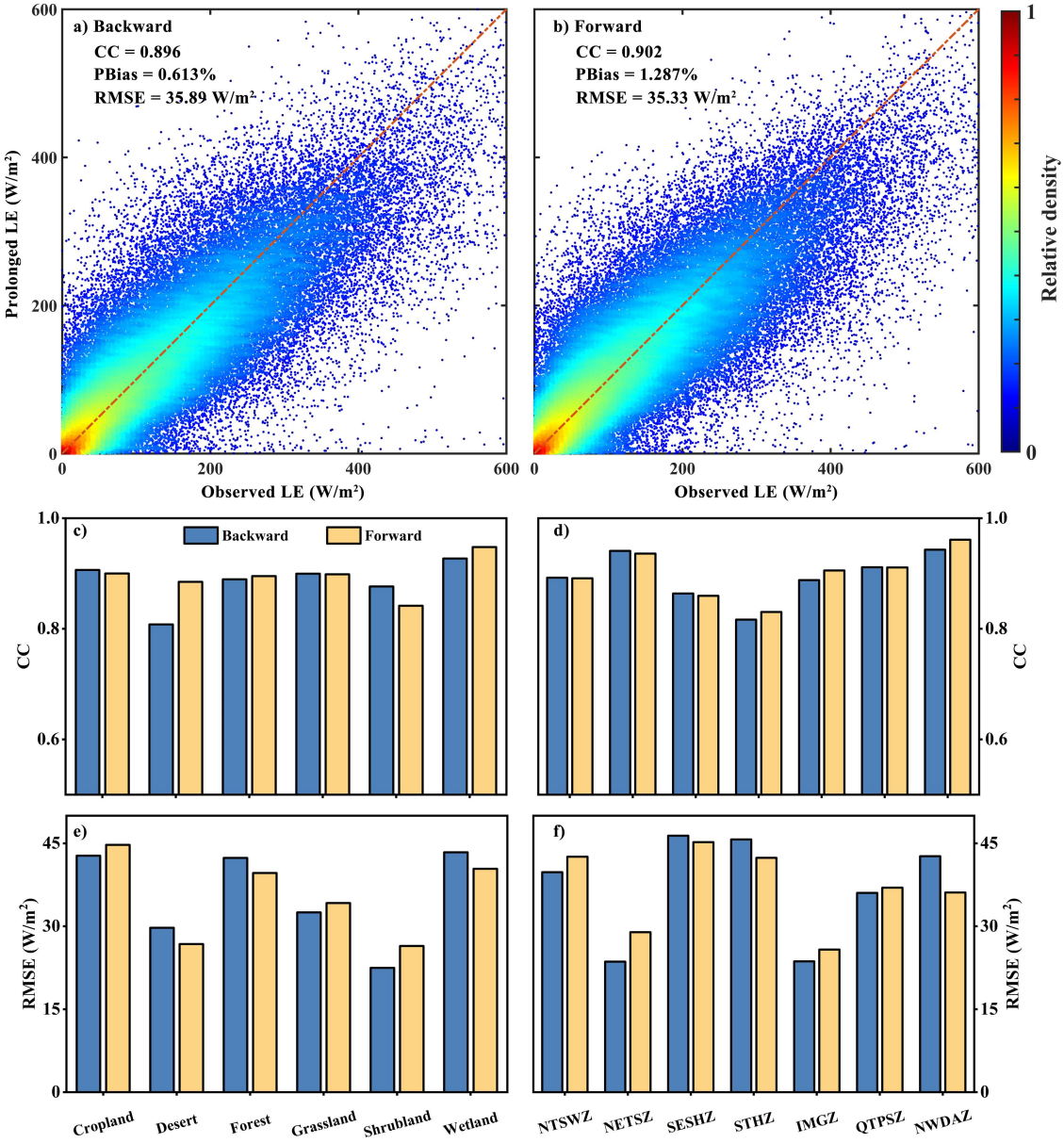

Figure 7: Consistency between backward (a) and forward (b) prolongation of half-hourly latent heat flux (LE), shown as scatter density plots of prolonged LE versus observed LE. The color scale represents the relative density of data points, normalized to highlight areas with higher concentration (warmer 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.

Detailed comments 35

Fig. 8a): from the figure description it is not clear whether the bars or the lines relate to the left or right y-axis. Please provide more information in the figure description.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we apologize for the ambiguity in the original figure description, which may have caused confusion regarding the correspondence between graphical elements and the y-axes in Fig. 8a. In the revised manuscript, we have clarified this explicitly in the figure caption. Specifically:

The bar plots represent RMSE values, corresponding to the left y-axis;

The line plots represent correlation coefficients (CC), corresponding to the right y-axis.

This clarification has been added directly to the figure caption to ensure that the relationship between visual elements and their respective axes is immediately clear to the reader.

We appreciate the reviewer's suggestion, which has helped improve the clarity and readability of the figure.

The revised ***title of Figure 8***:

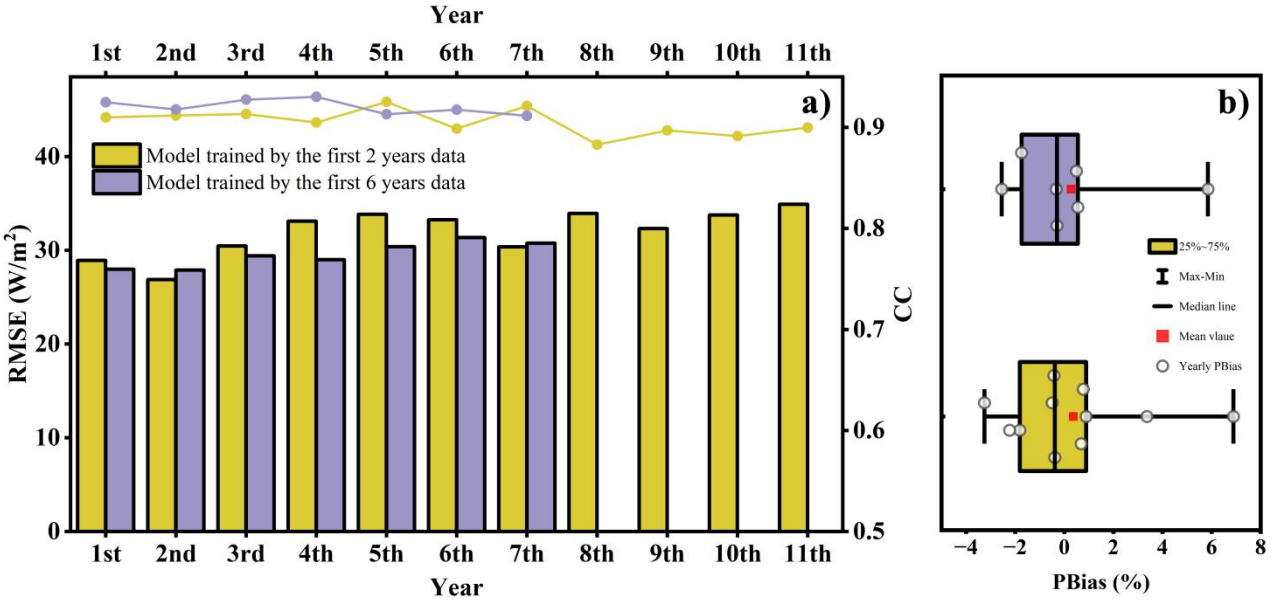

Figure 8: Temporal stability of forward prolongation performance for half-hourly LE using models trained with different lengths of observational data. **In panel (a), bars represent RMSE (left y-axis), while lines represent correlation coefficient (CC, right y-axis).**

Detailed comments 36

Fig. 9: might be removed or moved to the appendix. Instead give more details about statistics in the text in chap. 3.3.1

Fig. 10 and 11: same as for fig. 9, as daily and monthly values are just aggregates of half-hourly values in case less than 10% of missing data.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the presentation of multi-scale results can be streamlined and that additional emphasis on quantitative statistics would improve the clarity and conciseness of the manuscript.

In the revised version, we have made the following changes:

1. **Reorganization of figures:** Following the reviewer's suggestion, Fig. 10 and Fig. 11 (daily and monthly results) have been moved to the Appendix (now Fig. A1 and Fig. A2), as they represent aggregated forms of the half-hourly results and provide supporting rather than primary evidence.

2. **Structural simplification of the manuscript:** Sections 3.3.1, 3.3.2, and 3.3.3 have been merged into a single section, "3.3 Demonstration of different scale prolonged time series", to improve readability and avoid redundancy across temporal scales.

3. **Enhanced statistical description for Fig. 9:** In response to the reviewer's suggestion, we have substantially strengthened the statistical description in the text corresponding to Fig. 9, explicitly incorporating key performance metrics (CC, RMSE,

and PBIAS) for representative sites. This revision provides a clearer quantitative assessment of model performance at the half-hourly scale, rather than relying primarily on visual interpretation.

4. **Streamlined description of daily and monthly results:** The descriptions of daily and monthly scales (now Fig. A1 and Fig. A2) have been condensed and focused on key findings, avoiding repetitive explanations while retaining essential information.

Overall, these revisions improve the balance between visual presentation and quantitative analysis, enhance the clarity of the manuscript, and reduce unnecessary redundancy. We appreciate the reviewer's valuable suggestion, which has significantly improved the organization and readability of the paper.

The revised ***Section 3.3***:

3.3 Demonstration of different scale prolonged time series

To illustrate the performance of temporal prolongation at the half-hourly scale, Fig. 9 compares observed latent heat flux (LE) with AutoML-based prolonged series across representative sites covering diverse underlying surface types and climate zones. The prolonged series consistently reproduce the observed diurnal cycles and temporal variability, as reflected by high correlation coefficients (CC ranging from 0.737 to 0.982) and generally low RMSE values (approximately 4–45 W/m2), 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 robust capture of radiation-driven processes. 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 demonstrate that the proposed framework can reliably reproduce high-frequency LE dynamics at the half-hourly scale, providing a sound basis 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 that the model reliably captures evapotranspiration processes controlled by radiation conditions and vegetation seasonality. 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, suggesting that the model maintains robust temporal stability when transitioning across scales.

The revised ***title of Figure 9***:

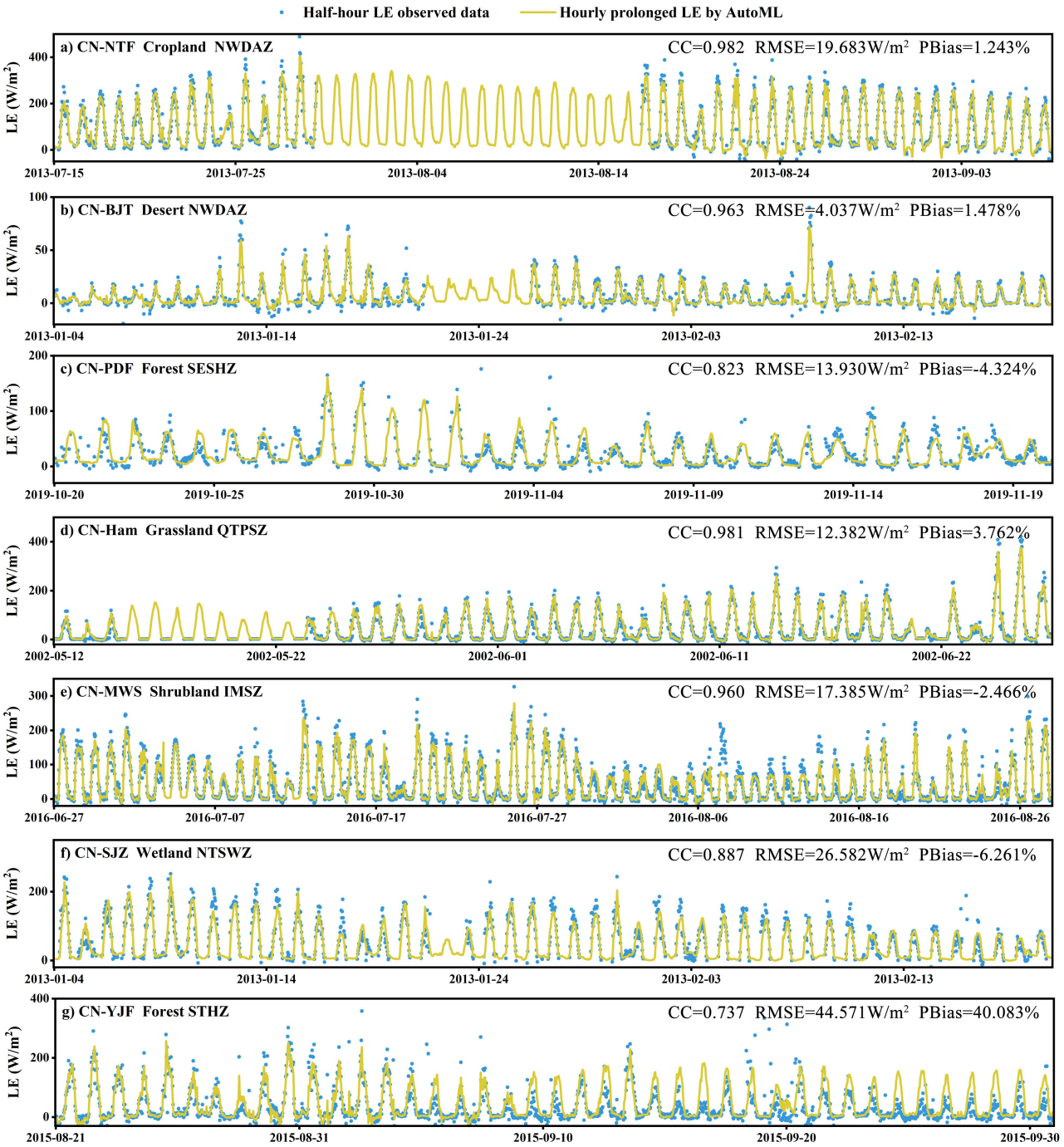

Figure 9: Demonstration of half-hourly prolonged LE time series across different typical sites.

Detailed comments 37

Fig. 12: 'l' is missing in word 'Daily' in c) and d)

Author Respond: Thank you for pointing out this typographical error. We have carefully checked the Figure and corrected the missing letter "l" in the word "Daily" in panels (c) and (d). The revised figure has been updated accordingly in the manuscript.

We appreciate the reviewer's attention to detail, which has helped improve the overall quality of the presentation.

The revised ***Figure 10***:

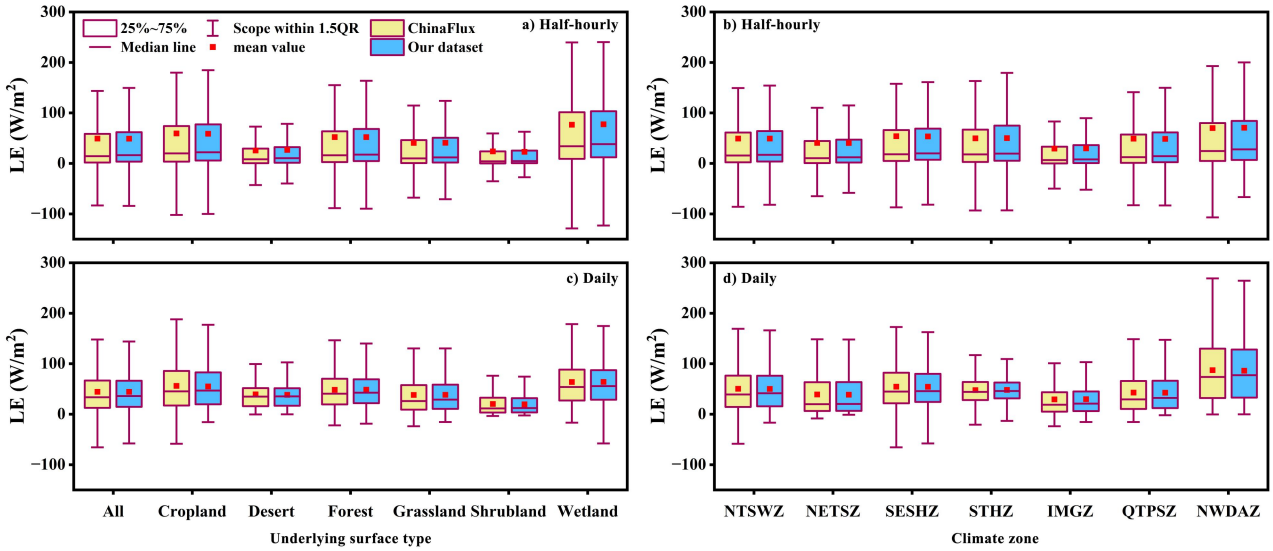

Figure 10: Comparison of data distributions between the reconstructed LE dataset and ChinaFlux observations at half-hourly (a,b) and daily (c,d) scales across different underlying surface types and climate zones.

Detailed comments 38

Table A1: make sure that words in the table header are not wrapped (looks ugly)

Line 669: What is meant with: '..that this station provides only interpolated data.'Are these the same 10 sites mentioned above with gap-filled data? Are those data treated as measured data?

Author Respond: Thank you for pointing out the issues and providing valuable suggestions.

1. Table formatting: We agree that the wrapped words in the table header reduced readability. In the revised manuscript, Table A1 has been reformatted to ensure that all header text is displayed without line wrapping, improving the visual clarity and presentation quality.

2. **Clarification of "interpolated data":** We apologize for the lack of clarity in the original description. The statement "The * in the station ID indicates that this station provides only interpolated data" refers to the same 10 sites mentioned earlier in the manuscript. These sites do not provide original observations after standard QA/QC filtering. Instead, they provide continuous time series within the observation period that already include gap-filled or interpolated values, without explicit distinction between measured and filled data.

3. **Treatment of these sites in this study:** In our analysis, the data from these 10 sites are treated consistently with those from the other 40 sites, i.e., all available values are considered as observational data (labeled as "T"). This treatment ensures consistency in model input across sites. Importantly, since the modeling is conducted independently for each site, the inclusion of these stations does not affect the results of other sites.

4. **Rationale for inclusion:** These sites are retained in the dataset to maximize spatial coverage and data availability, particularly in regions where flux observations are sparse. This provides users with a broader set of site-level information while maintaining methodological consistency.

We appreciate the reviewer's comments, which have helped us improve both the clarity of the table and the transparency of data usage in the manuscript.

The revised ***Table A1***:

Table A1. Basic information for 50 sites in China. **Stations marked with "*" provide only interpolated time series data (without QA/QC-filtered observations).**

Station ID	SITE Name	Longitude (°E)	Latitude (°N)	Climate zone	Underlying surface type	Elevation (m)	Annual average temperature (℃)	Annual precipitation (mm)	Number of LE Data	Missing ratio	Start yea	End year
1*	CN-JFS	107.1508	29.0217	SESHZ	Forest	2	9.5	631	35088	0.000	2020	2021
2*	CN-NQF	92.0167	31.6500	QTPSZ	Grassland	1409	4.6	255	87648	0.000	2014	2018
3*	CN-QYZ	115.0667	26.7333	SESHZ	Forest	111	17.9	1489	227904	0.000	2003	2015
4*	CN-XLD	112.4667	35.0167	NTSWZ	Forest	3500	1.5	747	35088	0.000	2016	2017
5*	CN-DHS	112.5343	23.1738	STHZ	Forest	1317	2.1	365	140256	0.000	2003	2011
6*	CN-HB3	101.3333	37.6667	QTPSZ	Grassland	145	5.6	420	140256	0.000	2003	2011
7*	CN-DXF	91.0833	30.8500	QTPSZ	Grassland	1411	15.2	830	122736	0.000	2004	2011
8*	CN-NMG	116.4040	43.3255	IMGZ	Grassland	1350	2.4	375	122736	0.000	2004	2011
9	CN-DSL	98.9406	38.8399	QTPSZ	Wetland	553	23.8	712	44427	0.000	2014	2016
10*	CN-YJF	102.1775	23.4739	STHZ	Grassland	4333	1.3	477	46800	0.001	2013	2016
11*	CN-BTM	111.9352	33.4997	SESHZ	Forest	1378	9.0	124	35040	0.001	2017	2019
12	CN-CLF	123.4703	44.5966	NETSZ	Cropland	3200	-1.7	580	43617	0.009	2018	2021
13	CN-Cng	123.5092	44.5934	NETSZ	Grassland	2400	21.5	1931	58312	0.023	2007	2010
14	CN-HaM	101.1800	37.3700	QTPSZ	Grassland	1080	3.8	800	50439	0.036	2002	2004
15	CN-MWS	107.2300	37.7100	IMSZ	Shrubland	15	12.2	12	83863	0.044	2012	2016
16	CN-HB1	101.3167	37.6167	QTPSZ	Grassland	143	6.4	500	100449	0.045	2004	2009
17	CN-LCA	114.6833	37.8833	NTSWZ	Cropland	876	9.8	37	66910	0.046	2013	2017
18	CN-XLH	116.6714	43.5544	IMGZ	Grassland	1350	2.4	380	162766	0.053	2006	2015
19	CN-JRF	119.2173	31.8068	SESHZ	Cropland	874	9.8	37	98580	0.059	2015	2020
20	CN-GCF	115.6667	39.1333	NTSWZ	Cropland	3480	1.2	720	43351	0.136	2020	2022
21	CN-YCA	116.5702	36.8290	NTSWZ	Cropland	3680	-1.8	580	120296	0.143	2003	2011
22	CN-REG	102.5500	32.8000	QTPSZ	Grassland	1556	7.3	185	83600	0.153	2015	2020
23	CN-DTH	113.0525	29.4875	SESHZ	Wetland	3850	-2.4	420	39638	0.197	2006	2008
24	CN-BN1	101.2653	21.9275	STHZ	Forest	1280	8.8	510	108316	0.228	2003	2011
25	CN-CBS	128.0958	42.4025	NETSZ	Forest	1530	8.3	293	99328	0.278	2003	2010
26	CN-JZF	121.2017	41.1480	NTSWZ	Cropland	1054	9.8	37	110654	0.369	2005	2014
27	CN-PJS	121.9646	40.9328	NTSWZ	Wetland	181	12.5	580	31866	0.393	2018	2020
28	CN-SJZ	118.9809	37.7664	NTSWZ	Wetland	328	12.5	580	84696	0.396	2011	2018
29	CN-Du2	116.2836	42.0466	IMGZ	Grassland	592	21.0	1490	39647	0.413	2015	2018
30	CN-HZF	121.0178	51.7811	NETSZ	Forest	1170	16.0	1432	49790	0.432	2014	2018
31	CN-Hgu	102.5900	32.8453	QTPSZ	Grassland	4520	-2.1	480	25881	0.438	2015	2017
32	CN-PDF	106.3167	26.6000	SESHZ	Forest	18	15.8	1090	46063	0.451	2015	2019
33	CN-LDF	101.1326	41.9993	NWDAZ	Desert	210	13.4	642	22885	0.473	2013	2015
34	CN-YS1	116.6563	40.4190	NTSWZ	Shrubland	3150	0.8	520	18426	0.473	2020	2021
35	CN-SJY	100.4800	34.3547	QTPSZ	Grassland	3439	-1.2	360	45943	0.476	2012	2016
36	CN-SSW	100.4933	38.7892	NWDAZ	Desert	1250	2.6	349	20520	0.486	2013	2015

37	CN-HB2	101.3119	37.6094	QTPSZ	Grassland	4	15.3	1022	53371	0.493	2015	2020
38	CN-NTF	101.1338	42.0048	NWDAZ	Cropland	1731	7.3	185	20381	0.493	2013	2015
39	CN-HYL	101.1239	41.9932	NWDAZ	Forest	875	9.8	37	21210	0.507	2013	2015
40	CN-ARF	100.4643	38.0473	QTPSZ	Forest	50	13.4	341	59548	0.515	2013	2019
41	CN-YS3	116.6588	40.4165	NTSWZ	Forest	3950	-2.9	531	16344	0.534	2020	2021
42	CN-ZYF	100.4464	38.9751	NWDAZ	Wetland	1460	7.3	185	75913	0.567	2013	2022
43	CN-BJT	100.3042	38.9150	NWDAZ	Desert	4	12.7	604	14415	0.571	2013	2014
44	CN-DMF	110.3315	41.6439	IMGZ	Grassland	756	21.5	1557	69214	0.605	2013	2022
45	CN-HZZ	100.3186	38.7652	NWDAZ	Desert	1502	3.8	280	66672	0.620	2013	2022
46	CN-YKF	100.2421	38.0142	QTPSZ	Grassland	144	25.0	1607	26064	0.628	2015	2018
47	CN-HHL	101.1335	41.9903	NWDAZ	Forest	4585	1.9	430	58114	0.647	2013	2022
48	CN-DMZ	100.3722	38.8555	NWDAZ	Cropland	3200	-1.7	550	38125	0.660	2013	2019
49	CN-HMF	100.9872	42.1135	NWDAZ	Desert	1594	7.3	185	34189	0.746	2015	2022
50	CN-QJF	118.2500	31.9667	SESHZ	Cropland	15	15.7	1080	12543	0.762	2017	2020