Response to reviewer #2 Comments - Round 1

Reviewer general comments: major comments. I congratulate the authors on creating a broad and valuable data set. Bringing together data from so many flux sites and applying common gap-filling methods is a valuable contribution to earth system science. While I applaud the effort, I do have a number of concerns.

Author Respond: We sincerely thank the reviewer for the thorough evaluation of our manuscript and for the many constructive comments and suggestions provided throughout the review process. We are particularly encouraged by the reviewer's recognition of the value of assembling and standardizing long-term observations from a large number of ChinaFlux sites into a unified benchmark dataset.

The reviewer's comments have significantly improved the manuscript. In particular, they helped us strengthen the documentation of data sources and methodologies, clarify the gap-filling and temporal prolongation workflow, improve the transparency and traceability of the released dataset, refine the interpretation of validation results, reduce unnecessary qualitative descriptions, streamline figures and discussions, and enhance the overall presentation of the data product. As a result, we believe that the revised manuscript is substantially improved in terms of scientific rigor, clarity, reproducibility, and compliance with the standards expected for an ESSD data publication.

We have made every effort to address all comments and concerns in a point-by-point manner. For each issue raised by the reviewer, we have either revised the manuscript accordingly or provided detailed explanations where clarification was needed. We are grateful for the reviewer's careful reading and insightful recommendations, which have greatly strengthened both the manuscript and the released dataset.

Once again, we sincerely appreciate the reviewer's time, expertise, and constructive feedback.

Major comments:

Major comments 1

The documentation of methods is not up to the standards required for publication.

Many of the data sets and data processing methods brought to bear in this data product are not cited. Citations for methods and data sets must be provided. If the methods are unique to this manuscript, they must be documented in this manuscript.

Examples of these methods and data set that must be cited in order for this manuscript to be suitable for publication include flux data processing and machine learning methods and the ERA-5 and MODIS data products.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the original manuscript did not provide sufficiently complete documentation and citations for several datasets, processing procedures, and methodological components used in the construction of the benchmark dataset.

Following the reviewer's suggestion, we have conducted a comprehensive revision of the manuscript. We systematically reviewed all datasets, preprocessing procedures, flux-data processing workflows, machine-learning methods, and auxiliary data sources used throughout the study. Appropriate references have now been added for the corresponding datasets and methodologies, including the ChinaFlux observations, flux-data processing procedures, AutoML and machine-learning algorithms, ERA5-Land reanalysis products, MODIS products, and other relevant components. In addition, where methodological descriptions were previously insufficient, we expanded the manuscript to provide clearer documentation of the procedures implemented in this study.

The specific revisions addressing each of these issues are described in detail in the corresponding point-by-point responses below. We believe these revisions substantially improve the transparency, reproducibility, and documentation quality of the manuscript and bring it more fully in line with the requirements of ESSD data publications.

Major comments 2

Many of the analyses presented, particularly the time series comparisons, are only semi-quantitative and accompanied by statements that aren't quantifiable or precise. While some examples of the data product are suitable for a manuscript documenting the data set, I find the time series figures and associated discussion to be excessive.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that several parts of the original manuscript relied on qualitative descriptions and that some of the time-series figures and accompanying discussions were more extensive than necessary for a data paper.

Following the reviewer's suggestion, we carefully revised the relevant sections throughout the manuscript. Specifically, we reviewed and modified qualitative expressions that were not sufficiently supported by quantitative evidence, replacing them with more objective and statistically interpretable descriptions wherever appropriate. We also streamlined the associated discussions to focus more directly on the evaluation and characteristics of the dataset.

In addition, several figures and analyses that were considered supplementary to the main objectives of the manuscript have been moved to the Appendix/Supplementary Materials. This revision reduces redundancy in the main text and improves the overall readability and conciseness of the manuscript while preserving the information for interested readers.

The detailed revisions addressing these issues are provided in the corresponding point-by-point responses below. We believe that these changes have substantially improved the clarity, rigor, and presentation of the manuscript and better align it with the expectations for an ESSD data paper.

Major comments 3

The methods that are used to fill and prolong the data are not clear. The methodology presents five different versions of AutoML but never explains what is finally done to create the filled data set or why this one version is chosen. This is critical to the end product.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the original description of the methodology was not sufficiently clear and could lead readers to believe that multiple AutoML-based gap-filling products were generated and that different AutoML configurations were used to produce the final released dataset. Following the reviewer's suggestion, we substantially revised the methodological description to clarify the complete production workflow and the role of each experiment.

Specifically, we now explicitly state that the four artificial gap scenarios (30 min, 1 d, 7 d, and 30 d) were designed solely for model evaluation and method comparison. Artificial gaps were introduced into valid observations to assess model performance under different missing-data durations, and these experiments were repeated multiple times to evaluate model robustness and stability. Likewise, the benchmark methods (MDS, RF, and XGBoost) were included only for comparative evaluation under identical testing conditions and were not used to generate the final released dataset.

After completion of the artificial-gap experiments, all artificial gaps were discarded. For the production dataset, each flux tower site was processed independently using a single site-specific final AutoML model trained with all available high-quality LE observations. This final model was then applied to fill the actual missing observations within the measurement period and subsequently used for temporal prolongation outside the observation period. Therefore, each site ultimately has only one final production model and one corresponding continuous LE time series.

To further improve clarity, we reorganized the relevant methodological descriptions, removed redundant explanations, and consolidated the complete workflow—including model training, validation, artificial-gap evaluation, benchmark comparison, final gap-filling, and temporal prolongation—into a single coherent description in Section 2.4.2.

We believe that these revisions clearly distinguish the model-evaluation framework from the final production procedure and make the generation process of the released dataset substantially more transparent. Detailed modifications are provided in the corresponding point-by-point responses below.

Major comments 4

Documentation for the eddy covariance (EC) flux sites is lacking, almost entirely, from the document. The EC flux sites all have instruments and data processing methods, in addition to site metadata (soils, vegetation, terrain) that are unknown in the present document. Many scientists have contributed to this data collection. Their methods must be documented.

The scientists who created these data should be credited for their contributions. While I admire the monumental effort of the assembling these data and this manuscript, I cannot support the publication of a data product based on undocumented original data sets.

Author Respond: We sincerely thank the reviewer for this important comment and fully agree that the original data contributors, site metadata, instrumentation, and processing methodologies underlying the ChinaFlux observations are essential components of a transparent and reusable benchmark dataset.

We would like to clarify that these supporting resources were already included in the data documentation through Appendix Table A2, although this was not sufficiently emphasized in the original manuscript. Specifically, Appendix Table A2 provides both data-access links and reference publication links for all ChinaFlux sites included in this study.

Through the data-access links, users can directly access the original site-level datasets maintained by the corresponding site teams and data repositories. Depending on site availability, these resources include not only latent heat flux observations but also additional measurements such as radiative fluxes, sensible heat fluxes, meteorological variables, atmospheric state variables, and soil observations.

More importantly, the reference publication links provide comprehensive documentation of the individual flux sites, including instrumentation, data processing procedures, observation protocols, ecosystem characteristics, vegetation information, terrain conditions, soil properties, and other site-specific information necessary for data interpretation and scientific applications. These publications also appropriately identify and credit the scientists and research teams responsible for establishing, maintaining, and processing the original observations.

Rather than reproducing highly heterogeneous site descriptions, instrumentation details, and processing workflows within the benchmark manuscript itself, we chose to provide direct links to the original datasets and site publications. We believe this approach offers several advantages: (1) it ensures access to the most complete and authoritative documentation available for each site; (2) it maintains full transparency and traceability of the underlying observations; and (3) it appropriately recognizes and credits the substantial contributions of the original site investigators and data providers whose long-term measurements form the foundation of this benchmark dataset.

To address the reviewer's concern and improve clarity, we have revised Section 5 and Appendix Table A2 to explicitly state that the provided links give access to both the original observational datasets and detailed site documentation, including instrumentation, processing methodologies, ecosystem characteristics, terrain information, and soil properties. We have also strengthened the description of these resources in the manuscript to better acknowledge the contributions of the original ChinaFlux site teams and data providers.

The detailed revisions are provided in the corresponding point-by-point responses below.

Major comments 5

Finally, I am concerned about the inclusion of investigator gap-filled data in what is fundamentally and gap-filling exercise. The document states that some site data were gap-filled by the EC flux investigators and that these data were retained as "true observations" out of necessity. I understand that this might be necessary but I am concerned that if these data are not flagged this will degrade the data set. I recommend that, if at all possible, the investigator gap-filled data should be excluded from the data set and this gap-filling exercise. Gap-filled data are not "true observations." The result is fitting a model to a model. If this is not possible, I suggest that, at a minimum, those sites that include investigator gap-filled data must be flagged so that a potential data product user would have the option to avoid using those data. Finally, I suggest that those sites with gap-filled data should be excluded from all analyses that are used to evaluate this gap-filling methodology.

Author Respond: We thank the reviewer for this thoughtful comment and fully agree that previously gap-filled data should be clearly distinguished from original observations in order to maintain the transparency, interpretability, and usability of the dataset.

Following the reviewer's suggestion, we carefully reviewed the treatment of these sites and clarified their role throughout the manuscript and dataset documentation.

First, although 10 of the selected sites provide only previously gap-filled LE time series from the original site investigators, these records are not treated as equivalent to original observations within the released dataset. For these sites, no additional AutoML gap-filling is performed within the observation period. Instead, the provided LE records are retained as supplied by the original data providers, and only temporal prolongation beyond the observation period is conducted. Consequently, these sites do not undergo the same within-period gap-filling procedure applied to sites with original observations.

Second, each site is modeled independently. Therefore, the 10 sites containing previously gap-filled LE data do not influence the model training, gap-filling performance, or temporal prolongation results of the remaining 40 sites that contain original flux observations. This design prevents the propagation of potential uncertainties from these sites into the broader benchmark dataset.

Third, in response to the reviewer's concern regarding transparency, all released data are explicitly labeled using the data-quality flag system. Specifically, T denotes original flux-tower-based LE records, F denotes LE values generated through AutoML gap-filling within the observation period, and P denotes LE values generated through temporal prolongation outside the observation period. In addition, the 10 sites containing previously gap-filled LE data are explicitly identified in Table A1 and in the accompanying Readme.txt file. This allows users to readily identify these sites and include or exclude them according to the requirements of their specific applications.

Finally, given the limited availability of long-term ChinaFlux observations, retaining these sites increases the spatial coverage and practical utility of the benchmark dataset. At the same time, the explicit documentation and site-level identification ensure that users who wish to work exclusively with original-observation sites can easily exclude these records from their analyses.

We have revised the manuscript, Table A1, and the dataset documentation to make these distinctions and data-quality indicators more explicit. Detailed modifications are provided in the corresponding point-by-point responses below.

Specific comments:

Detailed comments 1
Line 31. "SHAP-based". Please avoid using undefined acronyms in the abstract.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that acronyms should be defined when first introduced in the abstract to improve clarity and accessibility for a broad readership. Following your recommendation, we have replaced the acronym "SHAP" with its full name, "Shapley Additive Explanations (SHAP)", in the abstract.

The revised version in ***Abstract*** (**the Line 32 to 33** of the revised manuscript, the modified content is displayed in bold):

Interpretability analysis based on Shapley Additive Explanations (SHAP) indicates that energy supply consistently dominates LE variability, while vegetation state and water availability modulate their relative importance under different environmental conditions.

Detailed comments 2
Lines 53-54. "With the expansion of spatial and temporal scales" **Scales of what? And perhaps domain might be a better term for what you have in mind? I would also suggest that the term "scale" is overused and often ambiguous. Domain and resolution are often more precise terms.**

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the term "scale" can be ambiguous in this context. To improve clarity, we have revised the sentence to specify spatial and temporal domains and finer resolutions, which more precisely describe the expansion in both extent and granularity of the analyses.

The revised version in ***Introduction*** (**the Line 54 to 55** of the revised manuscript, the modified content is displayed in bold):

However, despite the central role of evapotranspiration in hydrological and ecosystem studies, existing observational approaches remain insufficient to fully meet the demands of regional- and long-term analyses. **With the expansion of spatial and temporal domains and finer resolutions**, land surface models, remote sensing retrievals, and data assimilation systems have become the primary means for generating spatially and temporally continuous ET estimates (Zhang et al., 2022; Yu et al., 2022; Zou et al., 2017).

Detailed comments 3
Line 67-68. "long periods of missing observations due to instrument maintenance, energy balance non-closure," **I don't think of energy balance non-closure being a source of missing data. It is a systematic problem with EC flux measurements that exists throughout EC flux measurement records.**

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that energy balance non-closure (EBC) is not a direct cause of missing observations. Rather, it is a well-recognized systematic issue associated with eddy covariance measurements that affects flux accuracy throughout the observational record. Following your suggestion, we have removed EBC from the list of factors causing data gaps and added a separate statement acknowledging it as a source of measurement uncertainty.

The revised version in ***Introduction*** (**the Line 69 to 73** of the revised manuscript, the modified content is displayed in bold):

However, EC flux time series are commonly affected by substantial data gaps and long periods of missing observations due to instrument maintenance, complex meteorological conditions, and data transmission interruptions, with large variations in effective observation lengths among sites. **In addition, EC measurements are subject to systematic uncertainties associated with energy balance non-closure, which may affect the accuracy of flux estimates even when observations are available.**

Detailed comments 4

Line 71-72. "Although the marginal distribution sampling (MDS) approach has been widely adopted as the official gap-filling algorithm"

My apologies, I was not aware of any "official" gap filling algorithm. I don't believe there is any regulatory agency identifying EC flux data methods. Can you please provide a citation for this method and its broad adoption? You address this later in the introduction. Perhaps these sections of the introduction should be merged.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that describing MDS as an "official" gap-filling algorithm was inaccurate, as no regulatory agency formally endorses any specific method. We have removed this phrasing from the earlier part of the introduction and instead provide a detailed discussion of MDS in the subsequent paragraph.

We also added clarification regarding the origin and implementation of MDS: the method was first proposed in 2005 by Reichstein et al., who systematically described it for flux data gap-filling. Since then, it has become one of the standard gap-filling methods for eddy covariance flux data. MDS identifies periods with similar meteorological conditions and fills missing fluxes using the mean or median of observed values, combining features of the lookup table (LUT) and mean diurnal variation (MDV) methods. It has been widely applied in international flux networks such as FLUXNET and CarboEurope.

To support its broad adoption, we cite several studies: Jia et al., 2015; Pastorello et al., 2020; Mahabbati et al., 2021; Vekuri et al., 2023, which demonstrate the extensive use of MDS in flux tower LE gap-filling and time-series reconstruction.

The revised version in ***Introduction*** **(the Line 73 to 80** 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, and long gaps lasting more than 30 d account for a considerable fraction of the missing records (Foltýnová et al., 2020; Zhu et al., 2022). **Although gap-filling is necessary to generate continuous LE time series, short gaps are often handled using various statistical or empirical approaches** Moreover, the overall observation durations of existing flux sites are relatively short, and conventional gap-filling methods do not support temporal prolongation, which further limits the applicability of flux observations in long-term studies.

The revised version in ***Introduction*** **(the Line 109 to 120** of the revised manuscript, the modified content is displayed in bold):

Although model-based and remote sensing approaches are widely used for regional-scale evapotranspiration estimation, gap-filling and time-series reconstruction of flux tower latent heat flux (LE) observations still primarily rely on statistical and empirical methods. **The marginal distribution sampling (MDS) method was first proposed and applied to flux data gap-filling by Reichstein et al. (2005), who systematically described the approach. Since then, it has become one of the standard gap-filling methods in eddy covariance flux data processing. Among these, MDS has been widely adopted as a standard approach due to its simplicity and reasonable performance for short gaps (Pastorello et al., 2020; Mahabbati et al., 2021; Vekuri et al., 2023). MDS fills missing flux data by identifying periods with similar meteorological conditions and using the mean or median of observed values, combining features of the lookup table (LUT) method and the mean diurnal variation (MDV) method. It has been widely applied in international flux networks such as FLUXNET**

and CarboEurope (Papale et al., 2006; Jia et al., 2015). However, because evapotranspiration is jointly controlled by energy availability, water supply, and vegetation dynamics, its responses are highly nonlinear and context dependent, which limits the ability of MDS—based on meteorological similarity assumptions—to represent complex causal relationships (Chen et al., 2012).

*The **Suplementary References***:

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.

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

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

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

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

Detailed comments 5

Lines 73-74. "it often fails to accurately reconstruct the magnitude and extremes of LE under long-gap conditions or during periods of extreme environmental variability."

Citation(s) please? I am not aware of this information or these studies.You address this later in the introduction. Perhaps these sections of the introduction should be merged.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the original statement was insufficiently supported by references and introduced methodological limitations before the relevant background discussion. Following your suggestion, we removed this statement from the earlier part of the Introduction to avoid redundancy and reserved the discussion of gap-filling method limitations for the subsequent paragraph where MDS is formally introduced.

In addition, we revised the description of MDS limitations to better reflect findings reported in the literature. Rather than stating that MDS cannot accurately reconstruct the magnitude and extremes of LE, we now emphasize that reconstruction uncertainty increases with gap length because the meteorological similarity assumptions underlying MDS may not fully represent the nonlinear interactions among environmental drivers controlling evapotranspiration. This interpretation is supported by previous evaluations of MDS performance (Moffat et al., 2007; Kang et al., 2019; Zhu et al., 2022).

The revised version in ***Introduction*** (**the Line 77 to 79** of the revised manuscript, the modified content is displayed in bold):

Although gap-filling is necessary to generate continuous LE time series, short gaps are often handled using various statistical or empirical approaches. **Nevertheless, accurately reconstructing long and continuous LE time series remains challenging because evapotranspiration is governed by complex interactions among atmospheric conditions, vegetation dynamics, and water availability.**

The revised version in ***Introduction*** (**the Line 121 to 125** of the revised manuscript, the modified content is displayed in bold):

However, because evapotranspiration is jointly controlled by energy availability, water supply, and vegetation dynamics, its responses are highly nonlinear and context dependent, which limits the ability of MDS—based on meteorological similarity assumptions—to represent complex causal relationships (Chen et al., 2012). Previous studies **have shown that the performance of MDS decreases as gap length increases, primarily because its meteorological similarity assumptions cannot fully capture the nonlinear interactions among environmental drivers controlling evapotranspiration. As a result, reconstruction uncertainty tends to increase for prolonged missing periods, particularly when flux variability is strongly influenced by changing vegetation and hydrological conditions** (Moffat et al., 2007; Kang et al., 2019; Zhu et al., 2022).

Detailed comments 6-7

Lines 74-75, "Moreover, the overall observation durations of existing flux sites are relatively short, and conventional gap-filling methods do not support temporal prolongation". This statement contradicts the earlier message about long gaps in the Fluxnet2015 data record. Please be consistent with the data and with your own message.

More significantly, I would not call extrapolating a set of measurements beyond their temporal extent with a model an observation.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the original wording could be misleading in two respects.

First, our intention was not to suggest that flux records are uniformly short, but rather that observation periods vary substantially among sites and that many sites do not provide sufficiently long continuous records for multi-decadal analyses. We have revised the text accordingly to avoid any inconsistency with our previous discussion of data gaps.

Second, we fully agree that model-based estimates generated beyond the observation period should not be described as observations. Our objective is not to extend observations themselves, but to reconstruct temporally continuous flux time series using observational constraints and environmental drivers. To avoid this ambiguity, we replaced the term "temporal prolongation" with a more precise description indicating extension of flux time series beyond the temporal coverage of available measurements.

The revised version in ***Introduction*** (**the Line 79 to 82** of the revised manuscript, the modified content is displayed in bold):

Moreover, **the available observation periods vary substantially among flux sites, and many sites do not provide sufficiently long records to support multi-decadal analyses. In addition, conventional gap-filling methods are primarily designed to reconstruct missing data within observation periods and are not intended for extending flux time series beyond the temporal coverage of available measurements, which further limits the applicability of flux datasets in long-term studies.**

Detailed comments 8

Lines 116-117. I suggest starting a new paragraph at some point where you introduce the idea of "data-driven" approaches. I would take care to distinguish this from the other methods you have cited since all of them use data. "Data-driven" alone does not distinguish among methods, in my opinion.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the term "data-driven" is potentially ambiguous because the previously discussed methods (e.g., MDS, lookup table, and

regression-based approaches) also rely on observational data. To improve clarity, we have started a new paragraph at this point and replaced “data-driven approaches” with the more specific term “machine learning-based approaches”. We also added a brief explanation highlighting the distinction between machine learning methods and traditional statistical or empirical approaches, namely their ability to learn complex nonlinear relationships directly from observations.

The revised version in ***Introduction*** (**the Line 130 to 135** of the revised manuscript, the modified content is displayed in bold):

Other conventional approaches, including mean diurnal variation, lookup table, and regression-based methods, exhibit similar limitations and are prone to systematic biases when data gaps are extensive or environmental conditions change markedly (Qian et al., 2024b).

Against this background, **machine learning-based approaches have emerged as a promising alternative for LE gap-filling and flux time-series reconstruction at flux tower sites. Unlike traditional statistical and empirical methods, machine learning approaches can learn complex nonlinear relationships between LE and its meteorological and surface drivers directly from observations.** These methods offer greater flexibility in handling long data gaps and complex environmental conditions when reference variables are continuously available (Kim et al., 2020; Zhu et al., 2022; Guo et al., 2022). Nevertheless, **most conventional machine learning models require manually specified model structures and hyperparameter tuning, while deep learning approaches are often sensitive to sample size and data distribution**, which can introduce instability when the number of available flux sites is limited (Khan et al., 2021).

Detailed comments 9

Line 123. “Automated machine learning (AutoML), by systematically exploring model architectures and optimizing hyperparameters,” Please include a citation or multiple citations describing this method. The following text explaining the benefits of this approach should also be expanded to include appropriate citations.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the original manuscript did not provide sufficient methodological background or references to support the description of AutoML and its advantages. Following your recommendation, we expanded this section and added several references describing the principles of AutoML, as well as the specific characteristics of the H2O AutoML framework used in this study.

Specifically, we now cite Hutter et al. (2019), which provides a comprehensive overview of automated machine learning methodologies, and LeDell and Poirier (2020), which describes the H2O AutoML framework and its capabilities for automated algorithm selection, hyperparameter optimization, and stacked ensemble construction. We also added references demonstrating the applicability of AutoML to environmental and geoscientific datasets characterized by nonlinear relationships and heterogeneous data sources.

The revised version in ***Introduction*** (**the Line 137 to 147** of the revised manuscript, the modified content is displayed in bold):

Automated machine learning (AutoML) **has emerged as an efficient framework for automatically selecting algorithms, optimizing hyperparameters, and generating ensemble models through systematic model exploration (Gijsbers et al., 2019; Hutter et al., 2019; LeDell and Poirier, 2020). In particular, the H2O AutoML platform integrates multiple machine learning algorithms and stacked ensemble strategies, enabling robust predictive performance while substantially reducing manual intervention in model development (LeDell and Poirier, 2020; LeDell et al., 2021; Madni et al., 2023). Previous studies have demonstrated the effectiveness of AutoML for environmental and geoscientific applications characterized by complex nonlinear relationships and heterogeneous datasets (Ferreira et al., 2021; Bhatnagar et al., 2022; Bodini, 2023; Xu et al., 2025). When combined with reanalysis data featuring high temporal and spatial continuity, AutoML provides a flexible framework for reconstructing continuous flux time series and estimating**

latent heat flux beyond observation periods using relationships learned from available measurements and environmental drivers, thereby laying a methodological foundation for constructing long-term, continuous, site-level evapotranspiration benchmark datasets.

The Suplementary References:

Gijsbers, P., LeDell, E., Thomas, J., Poirier, S., Bischl, B., and Vanschoren, J.: An Open Source AutoML Benchmark, arXiv [preprint], https://doi.org/10.48550/arXiv.1907.00909, 2019.
LeDell, E., 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.
LeDell, E., Poirier, S., et al.: H2O AutoML User Guide, H2O.ai, https://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html, 2021.

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

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

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

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

Detailed comments 10-12

Section 2.1

The entire paragraph introducing ChinaFlux has no citations. Many of the methods for data processing that are listed are published and the publications documenting those methods should be cited. Please add appropriate citations documenting this resource and its methods. Please add citations for the ChinaFlux data set(s).

If the network, its data sets and its methods are not published, this is a problem.

Line 161. "Previous studies have demonstrated…", Please cite at least a representative subset of those "previous studies."

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.

Finally, to ensure data accessibility, we have also provided reference links for each flux tower site, as shown in Table A2.

Based on your suggestion, we have revised the text to explicitly include these references and clarify the basis for the methodological statements. 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 172 to 195** 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.

In total, this study integrates and selects 50 ChinaFlux flux tower sites (Fig. 1a), covering six major underlying surface types in China, including grassland (16 sites), shrubland (2 sites), cropland (9 sites), forest (13 sites), desert (5 sites), and wetland (5 sites). These sites also span seven major climate zones: the Northeast Temperate Subhumid Zone (NETSZ, 4 sites), Inner Mongolia Temperate Semiarid Zone (IMSZ, 5 sites), Northern Temperate Subhumid Warm Zone (NTSWZ, 9 sites), Southeast Subtropical Humid Zone (SESHZ, 7 sites), Southern Tropical Humid Zone (STHZ, 3 sites), Northwest Desert Arid Zone (NWDAZ, 10 sites), and Qinghai–Tibet Plateau Semiarid Zone (QTPSZ, 12 sites). The selected sites exhibit substantial heterogeneity in climatic conditions, vegetation types, and hydrothermal regimes, enabling a robust representation of the spatial variability of evapotranspiration processes across China. Detailed information on site locations, underlying surface types, climate zone classifications, and observation periods is provided in Table A1**, and data source information for each flux towers can be seen in Table A2.**

*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.

*The **Suplementary Table A2***:

Table A2. Data Source Information for flux towers.

Station_ID	Reference DOI	Data DOI
1	10.11922/11-6035.csd.2023.0103.zh	http:/dx.doi.org/10.57760/sciencedb.o00119.00048
2	10.11922/11-6035.csd.2024.0098.zh	https://doi.org/10.57760/sciencedb.j00001.00821
3	10.11922/11-6035.csd.2024.0142.zh	https://doi.org/10.57760/sciencedb.ecodb.00105
4	10.11922/11-6035.csd.2023.0082.zh	https://doi.org/10.57760/sciencedb.ecodb.00095
5	10.11922/csdata.2020.0046.zh	http://www.dx.doi.org/10.11922/sciencedb.1009
6	10.11922/csdata.2020.0034.zh	http://www.sciencedb.cn/dataSet/handle/1007
7	10.11922/csdata.2020.0026.zh	http://www.dx.doi.org/10.11922/sciencedb.j00001.00248
8	https://doi.org/10.12199/nesdc.ecodb.chinaflux2003-20 10.2021.nmg.006	https://doi.org/10.12199/nesdc.ecodb.chinaflux2003-2010. 2021.nmg.006
9	10.11922/11-6035.csd.2024.0141.zh	https://www.scidb.cn/doi/10.11888/Atmos.tpdc.301233
10	10.11922/csdata.2020.0057.zh	http://www.dx.doi.org/10.11922/sciencedb.j00001.00066
11	10.11922/11-6035.csd.2023.0124.zh	http://www.dx.doi.org/10.57760/sciencedb.j00001.00859
12	10.11922/11-6035.csd.2023.0048.zh	https:/doi.org/10.57760/sciencedb.o00119.00068
13	https://doi.org/10.18140/FLX/1440209	https://fluxnet.org/sites/siteinfo/CN-Cng
14	https://doi.org/10.18140/FLX/1440190	https://fluxnet.org/sites/siteinfo/CN-HaM
15	10.17521/cjpe.2023.0001	https://doi.org/10.57760/sciencedb.ecodb.00094
16	10.11922/csdata.2020.0033.zh	http://www.sciencedb.cn/dataSet/handle/1010
17	10.11922/11-6035.csd.2023.0031.zh	https://doi.org/10.57760/sciencedb.o00119.00050

18	10.11922/11-6035.csd.2023.0059.zh	http://doi.org/10.57760/sciencedb.07140
19	10.11922/11-6035.csd.2023.0072.zh	http://dx.doi.org/10.57760/sciencedb.07131
20	10.11922/11-6035.csd.2023.0022.zh	https://doi.org/10.57760/sciencedb.o00119.00071
21	10.11922/csdata.2020.0044.zh	http://www.doi.org/10.11922/sciencedb.j00001.20002
22	10.11922/11-6035.csd.2023.0009.zh	https://dx.doi.org/10.57760/sciencedb.o00119.00052
23	10.11922/11-6035.csd.2023.0055.zh	https://doi.org/10.57760/sciencedb.o00119.00070
24	10.11922/csdata.2020.0037.zh	http://www.dx.doi.org/10.11922/sciencedb.j00001.00177
25	https://www.sciengine.com/CSD/doi/10.11922/csdata.2020.0041.zh	http://www.dx.doi.org/10.11922/sciencedb.1000
26	10.11922/11-6035.csd.2023.0007.zh	https://doi.org/10.57760/sciencedb.o00119.00062
27	10.11922/11-6035.csd.2023.0004.zh	https://doi.org/10.57760/sciencedb.06829
28	10.11922/11-6035.ncdc.2023.0007.zh	https://doi.org/10.57760/sciencedb.j00001.00867
29	https://doi.org/10.18140/FLX/1440140	https://fluxnet.org/sites/siteinfo/CN-Du2
30	10.11922/11-6035.csd.2023.0019.zh	https://doi.org/10.57760/sciencedb.o00119.00063
31	https://doi.org/10.18140/FLX/1669632	https://fluxnet.org/sites/siteinfo/CN-Hgu
32	10.11922/11-6035.csd.2023.0063.zh	https://doi.org/10.57760/sciencedb.o00119.00074
33	10.11922/11-6035.csd.2024.0099.zh	https://www.scidb.cn/doi/10.11888/Atmos.tpdc.301184
34	10.11922/11-6035.csd.2023.0030.zh	https://doi.org/10.57760/sciencedb.o00119.00035
35	10.11922/11-6035.csd.2023.0056.zh	https://doi.org/10.57760/sciencedb.o00119.00025
36	10.11922/11-6035.csd.2024.0099.zh	https://www.scidb.cn/doi/10.11888/Atmos.tpdc.301184
37	10.11922/11-6035.csd.2023.0012.zh	https://dx.doi.org/10.57760/sciencedb.06764
38	10.11922/11-6035.csd.2024.0099.zh	https://www.scidb.cn/doi/10.11888/Atmos.tpdc.301184
39	10.11922/11-6035.csd.2024.0099.zh	https://www.scidb.cn/doi/10.11888/Atmos.tpdc.301184
40	10.11922/11-6035.csd.2024.0141.zh	https://www.scidb.cn/doi/10.11888/Atmos.tpdc.301233
41	10.11922/11-6035.csd.2023.0030.zh	https://doi.org/10.57760/sciencedb.o00119.00035
42	10.11922/11-6035.csd.2024.0099.zh	https://www.scidb.cn/doi/10.11888/Atmos.tpdc.301184
43	10.11922/11-6035.csd.2024.0099.zh	https://www.scidb.cn/doi/10.11888/Atmos.tpdc.301184
44	10.11922/11-6035.csd.2023.0021.zh	https://doi.org/10.57760/sciencedb.o00119.00043
45	10.11922/11-6035.csd.2024.0099.zh	https://www.scidb.cn/doi/10.11888/Atmos.tpdc.301184
46	10.11922/11-6035.csd.2024.0141.zh	https://www.scidb.cn/doi/10.11888/Atmos.tpdc.301233
47	10.11922/11-6035.csd.2024.0099.zh	https://www.scidb.cn/doi/10.11888/Atmos.tpdc.301184
48	10.11922/11-6035.csd.2024.0099.zh	https://www.scidb.cn/doi/10.11888/Atmos.tpdc.301184
49	10.11922/11-6035.csd.2024.0099.zh	https://www.scidb.cn/doi/10.11888/Atmos.tpdc.301184
50	10.11922/11-6035.csd.2023.0001.zh	https://doi.org/10.57760/sciencedb.j00001.00779

Detailed comments 13-16

13. Figure 1a. The elevation scale should be eliminated. It is not relevant to the information presented about the sites (vegetation cover). It is colorful but misleading - it can confuse the reader regarding what is presented concerning site locations.

14. Figure 1b. This figure is difficult to understand. The same variable exists on 2 axes. More explanation of how to read this would be helpful.

15. Figure 1b. I believe this figure is saying that 15 sites have nearly zero observational gaps. In my experience with EC flux measurements this is nearly impossible. Rainfall and stable atmospheric conditions essentially always lead to some significant fraction (~20%?) of data loss. How is it possible to have sites with less than 5% missing data?

16. Figure 1d. Please include units on the x-axis.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we have carefully revised the figure and its caption to improve clarity and avoid potential misinterpretation.

First, as suggested, we removed the elevation color scale from Figure 1a. The purpose of panel (a) is to show the spatial distribution of sites across different vegetation and land-cover types. We agree that the elevation background was not directly relevant to this objective and could potentially distract readers or obscure the intended information. The revised panel now uses a cleaner background focused on land-cover classification and site locations.

Second, we agree that the original presentation of Figure 1b was insufficiently explained. To improve readability, we expanded the figure caption to explicitly describe the meaning of both graphical elements. The histogram represents the number of sites falling within different percentages of LE observation gaps, whereas the blue line shows the corresponding gap percentages for individual sites ordered by Site_ID. Additional clarification has been added to the caption to guide readers in interpreting the figure.

Third, regarding the unexpectedly low gap percentages for some sites, we appreciate this observation. Indeed, under normal eddy covariance processing workflows, it is uncommon for sites to exhibit extremely low percentages of missing data after quality control. In our dataset, this pattern is mainly attributable to ten sites that only provided interpolated LE time series rather than the original QA/QC-filtered observations. These sites are explicitly identified in Table A1 using a special marker. Because all sites were modeled independently, the use of these interpolated series does not affect model training or reconstruction results at other sites. To ensure transparency, we have further clarified this issue in the manuscript and maintained data-type flags in the released dataset so that users can distinguish between original observations and interpolated records. We believe retaining these sites provides additional benchmark information and broader spatial coverage while allowing users to make informed decisions regarding data selection.

Finally, following the reviewer's recommendation, we added the unit "years" to the x-axis label in Figure 1d.

*The **revised Figure 1***:

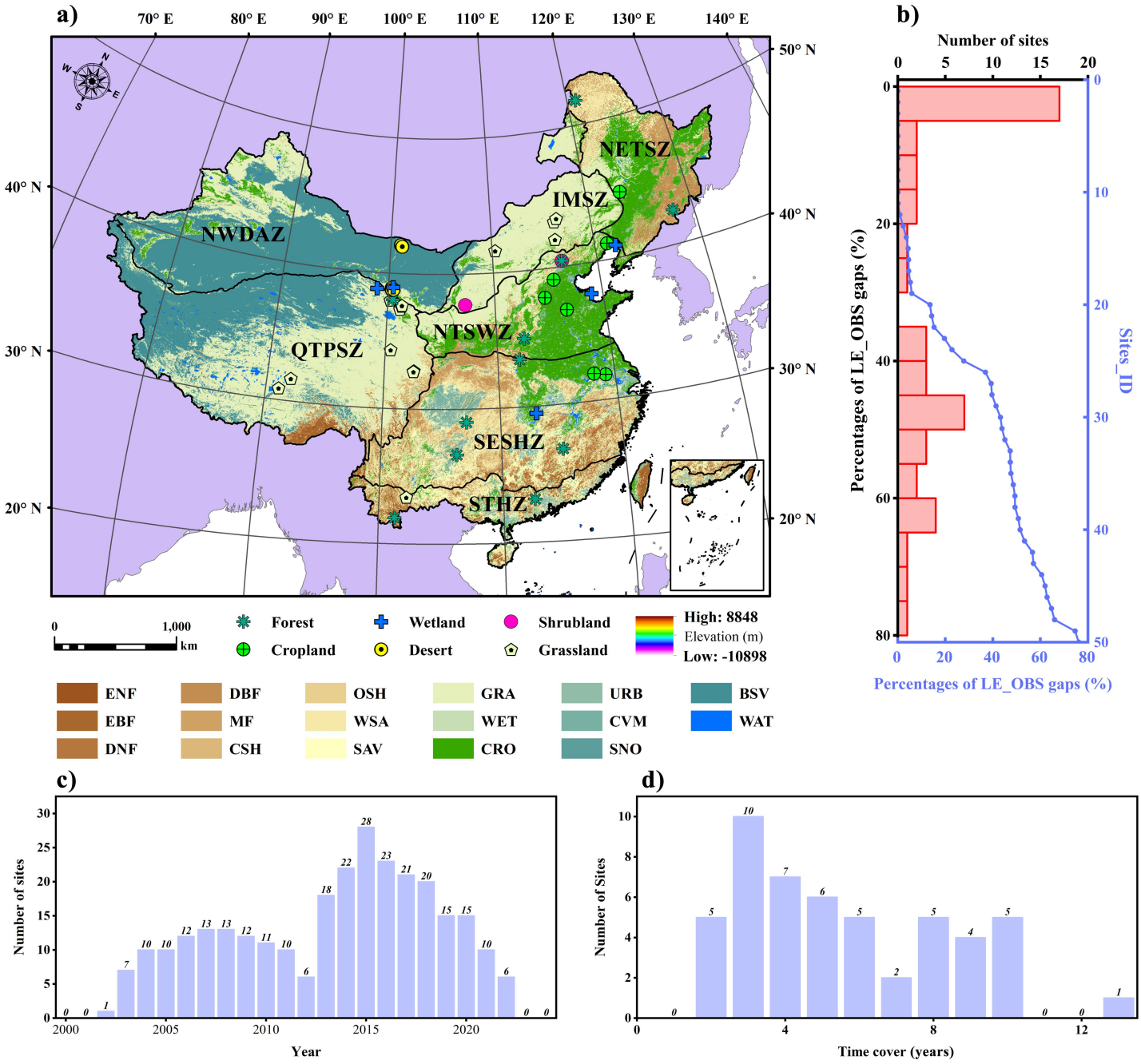

Figure 1: Spatial distribution of the selected ChinaFlux sites (a) and data coverage characteristics of half-hourly LE observations (b–d). **In panel (b), the histogram (left y-axis) represents the number of sites within different percentages of LE observation gaps (top x-axis), while the blue line shows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

Detailed comments 17-19

17. Lines 172-173. "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."

Does this mean that the gap filled data are included in the "observations?" If the data from these sites does not clarify what are observations vs. what is gap-filled that is a concern. Please clarify what methods have been followed here and what metadata are available for these sites. Ideally you are not gap-filling data that have already been gap-filled. I strongly recommend that you start only with observed fluxes before doing any gap filling.

18. I suggest that you do not include sites if they have only gap-filled flux records and you cannot identify the true observations.

19. I could accept the following compromise: If sites have gap filling that cannot be identified, please clearly identify these sites in your final data product. This would allow a data use to avoid these sites in their analyses.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we fully agree that it is critical to distinguish between original observations and already gap-filled data to maintain transparency and usability of the dataset.

First, although 10 of the selected sites only provide previously gap-filled LE time series, these series are not treated as "observations" in the dataset. For these sites, artificial gaps are introduced during training in the same manner as the other 40 sites, and no additional F-type gap-filling is performed within their observation periods. Only temporally prolonged data beyond the observation period are generated and explicitly flagged as P.

Each site is independently modeled, so these 10 sites do not affect the results or quality of the other 40 sites with original observations, ensuring high-quality data for model training and evaluation.

Second, in the final released dataset, each site is provided as a separate CSV file, and all data points are clearly labeled as T, F, or P: T denotes half-hourly LE derived from original flux tower observations; F represents gap-filled LE generated within the observation period using the AutoML framework; P indicates LE values produced through temporal prolongation beyond the observation period using the same framework. The 10 sites with only pre-gap-filled data are explicitly marked in Table A1 and in the Readme.txt file included with the dataset, allowing users to include or exclude these sites according to their specific research objectives.

Finally, given the overall scarcity of ChinaFlux data, retaining these 10 sites provides users with greater flexibility while fully documenting the nature of the data. Users who prefer to analyze only original observations can easily exclude these sites using the provided metadata and flags.

The revised version in ***Data and methodology*** (**the Line 199 to 211** 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 only provide continuous LE time series that have already been gap-filled within their observation periods. These gap-filled series are not considered "observations" in the dataset. For these 10 sites, artificial gaps are introduced in the same manner as for the other 40 sites to ensure a consistent training strategy, and only temporally prolonged data beyond the observation period are generated. Within the observation period, no gap-filled data are created for these sites. All sites are modeled independently, so the presence of these 10 sites does not affect the results or quality of the 40 original-observation sites. In the final dataset, each site corresponds to a separate CSV file, and all data are clearly classified and labeled to distinguish different data sources and processing stages. Detailed descriptions of the data flags and file structure are provided in Section 5 ("Data availability"). These 10 sites are explicitly identified in Table A1 and in the dataset's Readme.txt file, allowing users to include or exclude them according to their specific research needs.** Given the overall scarcity of ChinaFlux data, retaining these sites provides users with greater flexibility in selecting data for their specific research objectives.

Detailed comments 20-23

20. Appendix A: Table AI. I see that Table A1 has no citations and no investigators associated with these flux sites. If this data set is the only published record of these data, then perhaps this is appropriate. I am concerned, however, that each site should have metadata describing the site, its instrumentation and its data processing methods that is not available in this manuscript. Please include references where readers can find the details about these sites, their investigators and their methods. If these references do not exist, I am concerned about the value of this data set. Flux records with biological and environmental data limited to "forest" have limited value. It is true that many land surface characteristics can be gathered from satellite data sets but many cannot. If you do not have access to these data concerning the sites, please explain how this data set retains its value, and add this discussion to the manuscript.

21. Table A1: I am also very puzzled by the sites with a "missing ratio" of zero (as previously noted). How is this possible? What is the definition of the "missing ratio"?

22. Table A1: Several column headings have words that are broken across rows (e.g. Longitud e). Please correct this.

23. Table A1: Is elevation the altitude above sea level of the ground at the flux tower sites? What are the heights of the towers above ground?

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we thank the reviewer for these valuable comments regarding the metadata completeness, missing ratio definition, and site information presented in Appendix A (Table A1). Following the reviewer's suggestions, we have substantially revised the appendix and manuscript to improve transparency, data traceability, and clarity.

First, to address concerns regarding the availability of site metadata, instrumentation, data processing methods, and investigator information, we compiled the original references and data access links for all 50 flux sites and added them to a newly created Table A2 in the Appendix. Specifically, for each site, we now provide the corresponding publication DOI and/or data repository link, where readers can access detailed descriptions of site characteristics, instrumentation configurations, data processing procedures, principal investigators, and associated research methodologies. These additions not only acknowledge the contributions of the original site investigators but also substantially improve the transparency, reproducibility, and scientific utility of the dataset.

Second, we agree that classifying all forest sites simply as "Forest" provides limited ecological information. Therefore, we further refined the land-cover descriptions of all forest flux sites and incorporated these classifications into Table A1. The forest subcategories are as follows:

CN-BN1: Tropical Seasonal Rainforest

CN-BTM: Natural Oak Forest

CN-CBS: Broad-Leaved Red Pine Forest

CN-DHS: Mixed Coniferous Broad-Leaved Forest

CN-HHL: Poplar–Tamarisk Mixed Forest

CN-HYL: Euphrates Poplar Forest

CN-HZF: Temperate Boreal Coniferous Forest

CN-JFS: Subtropical Evergreen Broad-Leaved Forest

CN-PDF: Shrub-Dominated Secondary Forest

CN-QYZ: Coniferous Plantation

CN-XLD: Quercus Variabilis Plantation

CN-YS3: Coniferous Plantation

Accordingly, the caption of Table A1 has been revised to better describe the site information included in the table.

Third, regarding the sites with a missing ratio equal to zero, we apologize for the insufficient explanation in the original manuscript. As noted in our previous response, these correspond to the ten sites marked with an asterisk (*), for which only gap-filled latent heat flux (LE) time series were publicly available, while the original QA/QC-filtered observations were not provided. Since the released datasets contain complete gap-filled records without missing half-hourly values, their missing ratio is reported as zero.

In addition, we now explicitly define the missing ratio as:

Missing Ratio = Number of missing half-hourly observations/Total number of half-hourly observations during the observation period

Fourth, we appreciate the reviewer's observation regarding formatting issues in Table A1. The column headings have been reformatted, and all words previously split across multiple lines (e.g., "Longitude", "Latitude") have been corrected to improve readability and presentation quality.

Finally, we clarify that the "Elevation" column in Table A1 refers to the ground elevation of each flux tower site above mean sea level. The tower heights above ground have now been added in parentheses in Table A1 for all sites where this information is available. This ensures that both site elevation and tower height are clearly documented for readers.

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). **Elevation refers to the ground elevation of the flux tower site above mean sea level. Height refers to the heights of the towers above ground (the value in parentheses). The forest sites are subdivided as follows, CN-BN1: Tropical Seasonal Rainforest; CN-BTM: Natural Oak Forest; CN-CBS: Broad-Leaved Red Pine Forest; CN-DHS: Mixed Coniferous Broad-Leaved Forest; CN-HHL: Poplar–Tamarisk Mixed Forest; CN-HYL: Euphrates Poplar Forest; CN-HZF: Temperate Boreal Coniferous Forest; CN-JFS: Subtropical Evergreen Broad-Leaved Forest; CN-PDF: Shrub-Dominated Secondary Forest; CN-QYZ: Coniferous Plantation; CN-XLD: Quercus variabilis Plantation; CN-YS3: Coniferous Plantation.**

Station ID	SITE Name	Longitude (°E)	Latitude (°N)	Climate zone	Underlying surface type	Elevation and Height (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	1543 (28)	8.2	1434	35088	0.000	2020	2021
2*	CN-NQF	92.0167	31.6500	QTPSZ	Grassland	4585 (2.3)	1.9	430	87648	0.000	2014	2018
3*	CN-QYZ	115.0667	26.7333	SESHZ	Forest	110 (39.6)	17.9	1489	227904	0.000	2003	2015
4*	CN-XLD	112.4667	35.0167	NTSWZ	Forest	410 (36)	13.4	641.7	35088	0.000	2016	2017
5*	CN-DHS	112.5343	23.1738	STHZ	Forest	300 (38)	22.5	1714	140256	0.000	2003	2011
6*	CN-HB3	101.3333	37.6667	QTPSZ	Grassland	3200 (2.5)	-1.7	580	140256	0.000	2003	2011
7*	CN-DXF	91.0833	30.8500	QTPSZ	Grassland	4333 (2.2)	1.3	477	122736	0.000	2004	2011

8*	CN-NMG	116.4040	43.3255	IMGZ	Grassland	1317 (2.5)	2.1	365	122736	0.000	2004	2011
9	CN-DSL	98.9406	38.8399	QTPSZ	Wetland	3439 (4.5)	-1.2	410	44427	0.001	2014	2016
10*	CN-YJF	102.1775	23.4739	STHZ	Grassland	553 (13.9)	23.8	711.8	46800	0.001	2013	2016
11*	CN-BTM	111.9352	33.4997	SESHZ	Forest	1410.7 (38)	15.2	830	35040	0.001	2017	2019
12	CN-CLF	123.4703	44.5966	NETSZ	Cropland	143 (2)	6.4	500	43617	0.009	2018	2021
13	CN-Cng	123.5092	44.5934	NETSZ	Grassland	145 (2)	5.6	420	58312	0.023	2007	2010
14	CN-HaM	101.1800	37.3700	QTPSZ	Grassland	3150 (2.2)	0.8	520	50439	0.036	2002	2004
15	CN-MWS	107.2300	37.7100	IMSZ	Shrubland	1530 (6.2)	8.3	293	83863	0.044	2012	2016
16	CN-HB1	101.3167	37.6167	QTPSZ	Grassland	3200 (2.5)	-1.7	580	100449	0.045	2004	2009
17	CN-LCA	114.6833	37.8833	NTSWZ	Cropland	50.1 (3.5)	13.4	341	66910	0.046	2013	2017
18	CN-XLH	116.6714	43.5544	IMGZ	Grassland	1250 (4)	2.6	349	162766	0.053	2006	2015
19	CN-JRF	119.2173	31.8068	SESHZ	Cropland	15 (3.5)	15.7	1080	98580	0.059	2015	2020
20	CN-GCF	115.6667	39.1333	NTSWZ	Cropland	15 (4.5)	12.2	528	43351	0.136	2020	2022
21	CN-YCA	116.5702	36.8290	NTSWZ	Cropland	28 (3.3)	13.1	528	120296	0.143	2003	2011
22	CN-REG	102.5500	32.8000	QTPSZ	Grassland	3500 (2.5)	1.5	747	83600	0.153	2015	2020
23	CN-DTH	113.0525	29.4875	SESHZ	Wetland	68 (6.5)	16.4	1382	39638	0.197	2006	2008
24	CN-BN1	101.2653	21.9275	STHZ	Forest	756 (72)	21.5	1556.8	108316	0.228	2003	2011
25	CN-CBS	128.0958	42.4025	NETSZ	Forest	1080 (61.5)	3.8	800	99328	0.278	2003	2010
26	CN-JZF	121.2017	41.1480	NTSWZ	Cropland	23 (5)	8.5	640	110654	0.369	2005	2014
27	CN-PJS	121.9646	40.9328	NTSWZ	Wetland	3.8 (4.2)	9.5	631	31866	0.393	2018	2020
28	CN-SJZ	118.9809	37.7664	NTSWZ	Wetland	3.5 (3)	12.66	604	84696	0.396	2011	2018
29	CN-Du2	116.2836	42.0466	IMGZ	Grassland	1350 (5)	2.4	380	39647	0.413	2015	2018
30	CN-HZF	121.0178	51.7811	NETSZ	Forest	773 (37)	-4.4	481.6	49790	0.432	2014	2018
31	CN-Hgu	102.5900	32.8453	QTPSZ	Grassland	3480 (2.5)	1.2	720	25881	0.438	2015	2017
32	CN-PDF	106.3167	26.6000	SESHZ	Forest	1170 (24)	15.96	1432	46063	0.451	2015	2019
33	CN-LDF	101.1326	41.9993	NWDAZ	Desert	878 (3)	9.75	37.31	22885	0.473	2013	2015
34	CN-YS1	116.6563	40.4190	NTSWZ	Shrubland	141 (30)	12.5	580	18426	0.473	2020	2021
35	CN-SJY	100.4800	34.3547	QTPSZ	Grassland	3950 (2.2)	-2.9	531	45943	0.476	2012	2016
36	CN-SSW	100.4933	38.7892	NWDAZ	Desert	1594 (4.6)	7.29	184.83	20520	0.486	2013	2015
37	CN-HB2	101.3119	37.6094	QTPSZ	Grassland	3200 (2.2)	-1.7	550	53371	0.493	2015	2020
38	CN-NTF	101.1338	42.0048	NWDAZ	Cropland	875 (3.5)	9.75	37.31	20381	0.493	2013	2015
39	CN-HYL	101.1239	41.9932	NWDAZ	Forest	876 (22)	9.75	37.31	21210	0.507	2013	2015
40	CN-ARF	100.4643	38.0473	QTPSZ	Grassland	3033 (3.5)	-0.8	410	59548	0.515	2013	2019
41	CN-YS3	116.6588	40.4165	NTSWZ	Forest	328 (30)	12.5	580	16344	0.534	2020	2021
42	CN-ZYF	100.4464	38.9751	NWDAZ	Wetland	1460 (5.2)	7.29	184.83	75913	0.567	2013	2022
43	CN-BJT	100.3042	38.9150	NWDAZ	Desert	1562 (4.6)	7.29	184.83	14415	0.571	2013	2014
44	CN-DMF	110.3315	41.6439	IMGZ	Grassland	1409 (2.3)	4.6	255.2	69214	0.605	2013	2022
45	CN-HZZ	100.3186	38.7652	NWDAZ	Desert	1731 (4.6)	7.29	184.83	66672	0.620	2013	2022
46	CN-YKF	100.2421	38.0142	QTPSZ	Grassland	4146 (3.2)	2.3	41.0	26064	0.628	2015	2018
47	CN-HHL	101.1335	41.9903	NWDAZ	Forest	874 (22)	9.75	37.31	58114	0.647	2013	2022
48	CN-DMZ	100.3722	38.8555	NWDAZ	Cropland	1556 (4.5)	7.29	184.83	38125	0.660	2013	2019
49	CN-HMF	100.9872	42.1135	NWDAZ	Desert	1054 (4.5)	9.75	37.31	34189	0.746	2015	2022
50	CN-QJF	118.2500	31.9667	SESHZ	Cropland	18 (1.5)	15.8	1090	12543	0.762	2017	2020

The revised ***Table A2*** can be seen in ***"Detailed comments 10-12"***.

Detailed comments 24

Section 2.2.

Please include citations for the ERA-5 system and data sets. Web links are not sufficient documentation for these products. The lack of citations hinders readers from reviewing the details of the data products, limited the reproducibility and traceability of this work, and does not acknowledge the years of effort needed to make these data products publicly accessible.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we fully agree that formal citations should be provided for the ERA5 and ERA5-Land products rather than relying solely on web links. Proper citation not only improves reproducibility and traceability but also appropriately acknowledges the substantial effort involved in developing and maintaining these widely used reanalysis datasets.

Following your recommendation, we have added citations to both the ERA5 global reanalysis description paper (Hersbach et al., 2020), (Copernicus Climate Change Service, 2022), and the ERA5-Land dataset paper (Muñoz-Sabater et al., 2021) in Section 2.2. These references provide comprehensive descriptions of the data assimilation system, land-surface modeling framework, dataset characteristics, and validation procedures underlying the products used in this study.

The revised version in ***Data and methodology*** **(the Line 229 to 244** 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 **(Hersbach et al., 2020; Muñoz-Sabater et al., 2021)**. ERA5-Land is produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) and provides globally continuous coverage with high spatiotemporal consistency, offering a uniform background for flux reconstruction across different climate zones and land surface conditions **(Muñoz-Sabater et al., 2021; Copernicus Climate Change Service, 2022)**. Hourly ERA5-Land data were extracted at each site location from the ECMWF/ERA5_LAND/HOURLY dataset using Google Earth Engine (GEE; https://code.earthengine.google.com/, last access: 11 November 2025). Bilinear interpolation was applied to derive representative values at the site scale, and all timestamps were converted from Coordinated Universal Time (UTC) to local time to ensure consistency with flux tower observations. To match the half-hourly temporal resolution of the EC measurements, hourly ERA5-Land variables were resampled to a half-hourly scale using linear interpolation. The ERA5-Land variables used in this study include latent heat flux (LE), surface air pressure (PA), precipitation (Pre), relative humidity (RH), surface net solar radiation (denoted here as Rn, following the ERA5-Land variable definition), downwards solar radiation (Rs), surface runoff (Runoff), volumetric soil water in the 0–7 cm layer (SMC_1) and volumetric soil water in the 7–28 cm layer (SMC_2), air temperature (Temp), vapor pressure deficit (VPD), and 10 m wind speed (WS). These variables collectively characterize key controls on evapotranspiration, including surface energy balance, atmospheric moisture and aerodynamic conditions, and soil moisture constraints.

The ***Suplementary References***:

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

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

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 25

Section 2.3.

Please include citations for the MODIS data products that are being used here. As with ERA-5, web links are not sufficient documentation. The lack of citations hinders readers from reviewing the details of the data products, limited the reproducibility and traceability of this work, and does not acknowledge the years of effort needed to make these data products publicly accessible.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we fully agree that formal citations should be provided for the MODIS products used in this study rather than relying solely on web links. Proper citation improves reproducibility and traceability while appropriately acknowledging the substantial effort invested in the development, validation, maintenance, and public distribution of these products.

Following your recommendation, we added citations for both the MODIS vegetation index product (MOD13Q1) and the MODIS leaf area index product (MOD15A2H), including the official dataset references, product documentation, and key algorithm description papers. Specifically, we cite Didan (2015) and Didan and Barreto Munoz (2019) for the MOD13Q1 vegetation index product, and Myneni et al. (2015), MODIS LAI/FPAR Product Team (2020), and Myneni et al. (2002) for the MOD15A2H LAI product.

These references provide comprehensive information regarding product generation algorithms, quality assessment procedures, spatial and temporal characteristics, and recommended usage, thereby improving the transparency and reproducibility of the dataset construction process.

The revised version in ***Data and methodology*** (**the Line 245 to 261** of the revised manuscript, the modified content is displayed in bold):

2.3 MODIS data

To complement the meteorological drivers and better represent vegetation conditions, remotely sensed products from the Moderate Resolution Imaging Spectroradiometer (MODIS) were used to derive the normalized difference vegetation index (NDVI) **(Didan, 2015)** and leaf area index (LAI) **(Myneni et al., 2015)**. NDVI data were obtained from the MODIS MOD13Q1 (Terra) and MYD13Q1 (Aqua) products, while LAI data were derived from the MODIS MOD15A2H and MYD15A2H products. All products were converted to physical values using the corresponding scale factors provided in the official product user guides, with NDVI scaled by 0.0001 **(Didan and Barreto Munoz, 2019)** and LAI scaled by 0.1° **(MODIS LAI/FPAR Product Team, 2020)**. NDVI and LAI data were extracted at the site scale using GEE (https://code.earthengine.google.com/, last access: 11 November 2025), with bilinear interpolation applied for spatial sampling. The spatial resolutions of the NDVI and LAI products are 250 m and 500 m, respectively. Given the relatively low temporal resolution of MODIS products **(i.e., 16 day composites for NDVI and 8-day composites for LAI)**, NDVI and LAI were assumed to remain constant within each compositing period, and the corresponding values were assigned uniformly to all

half-hourly (and daily) time steps within that period to ensure temporal alignment with the half-hourly LE time series. **This assumption is commonly adopted in flux–remote sensing integration studies and is generally considered reasonable because MODIS compositing procedures are designed to reduce short-term noise (e.g., cloud contamination and atmospheric effects) and to represent average vegetation conditions over the compositing interval (Didan et al., 2015; Myneni et al., 2002).**

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.

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.

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.

Detailed comments 26
Lines 221-222. "Given that the AutoML framework is primarily based on tree-based algorithms…" **The AutoML code should be documented and accessible.**

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we fully agree that the code used to generate the dataset should be publicly accessible to ensure transparency, reproducibility, and long-term usability of the data product.

To address this, we have made the complete AutoML code used in this study publicly available through Zenodo with a permanent DOI: Qian et al. (2026): https://doi.org/10.5281/zenodo.18194590.

The repository contains the source code, workflow descriptions, and documentation required to reproduce the model training, gap-filling, and temporal prolongation procedures presented in this study.

In addition, detailed information regarding code availability has been included in Section 5 ("Data availability"), where both the dataset and the associated AutoML code repository are explicitly provided.

We believe that the public release of the code substantially improves the reproducibility and transparency of the dataset and enables users to fully examine, reproduce, and extend the methodology presented in this work.

Detailed comments 27

27. Section 2.4.

As with the previous two sections, no citations for the methodology are presented. For example, lines 229-231 note that data were quality screened but no details of the methods are presented and no citation for existing methods are given. This is not sufficient documentation for a data set. This is not at all reproducible.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the original version of Section 2.4 did not provide sufficient methodological references to support the data quality screening and model evaluation procedures.

Following your suggestion, we have substantially improved the documentation of the methodology by adding citations to the established processing frameworks used in this study. Specifically, the quality screening procedure is now explicitly linked to the standard processing workflows of FLUXNET and ChinaFlux through citations to Pastorello et al. (2020), Yu et al. (2006), and Zheng et al. (2019). These references provide detailed descriptions of eddy covariance quality-control procedures, including QA/QC flagging, turbulence filtering, and data screening protocols.

In addition, we added a citation to "Moffat et al. (2007) and Li et al. (2025)" to document the artificial-gap evaluation strategy and the use of benchmark gap-filling methods for performance assessment. These study are widely recognized as a foundational reference for evaluating gap-filling approaches using artificially generated gaps and for comparing different gap-filling methodologies.

These additions improve the transparency, traceability, and reproducibility of the dataset generation process by clearly linking all major methodological components to established and publicly available references.

The revised version in ***Data and methodology*** (**the Line 271 to 284** of the revised manuscript, the modified content is displayed in bold):

2.4 Generation of gap-filling

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

*The **Suplementary References**:*

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.

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.

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.

Moffat, A. M., Papale, D., Reichstein, M., Hollinger, D. Y., Richardson, A. D., Barr, A. G., Beckstein, C., Braswell, B. H., Churkina, G., Desai, A. R., Falge, E., Gove, J. H., Heimann, M., Hui, D., Jarvis, A. J., Kattge, J., Noormets, A., and Stauch, V. J.: Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes, Agricultural and Forest Meteorology, 147, 209–232, https://doi.org/10.1016/j.agrformet.2007.08.011, 2007.

Detailed comments 28 and 30

28. The flowchart, Figure 2, is very helpful. I hope that more details (e.g. "Determine optimal model") are described later in the text.

30. Lines 242-244. "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." What defines optimal? This detail must be explained. There are many metrics that can be used to define optimality.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the original description of "optimal model" was insufficient because it did not explicitly state which metric was used to define model optimality. We have revised Section 2.4.1 to provide a clearer and more reproducible description of the H2O AutoML modeling procedure and model selection criterion. And we have also added relevant references.

In the revised manuscript, we clarify that the optimal model was selected automatically by the H2O AutoML framework based on the root mean square error (RMSE) calculated on the validation set. Specifically, candidate models generated during the AutoML search process were ranked in the H2O AutoML leaderboard, and the leader model with the lowest validation RMSE was selected as the final optimal model for each flux tower site. This leader model could be an individual machine learning algorithm, such as GBM, DRF, XGBoost, GLM, or DNN, or a stacked ensemble model, depending on the site-specific validation results.

We also expanded the methodological description of H2O AutoML by specifying the candidate algorithms included in the search space, the use of random grid search for exploring model configurations and hyperparameter combinations, and the role of cross-validation and validation performance in model ranking.

These revisions provide a clear definition of model optimality and improve the transparency and reproducibility of the model selection procedure.

The revised version in ***Data and methodology*** (**the Line 286 to 311** of the revised manuscript, the modified content is displayed in bold):

2.4.1 AutoML

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

During the AutoML search process, multiple commonly used regression algorithms, including tree-based models and their ensemble variants, were evaluated, and the optimal model was automatically selected based on validation performance. **Specifically, the H2O AutoML framework evaluates a suite of candidate algorithms, including gradient boosting machine (GBM), distributed random forest (DRF), extreme gradient boosting (XGBoost), generalized linear models (GLM), and deep learning neural networks (DNN), as well as stacked ensemble models that combine multiple base learners. During the automated search process, multiple model configurations and hyperparameter combinations are explored through a random grid search strategy, and models are ranked based on cross-validation and validation performance. In this study, root mean square error (RMSE) on the validation set was used as the criterion for defining model optimality. The final model for each site was selected as the leader model from the H2O AutoML leaderboard, corresponding to the candidate model with the lowest validation RMSE among all evaluated algorithms and hyperparameter configurations. The final model for each site was selected as the leader model from the AutoML leaderboard, which represents the best-performing model (including possible ensemble models) under the given validation metrics. 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 29

The QC flags appear to differential Filled data from Prolonged data. This is very helpful. I do not support, however, calling investigator-filled data "True observational data." This point remains ambiguous. Please clarify. If investigator-filled data cannot be identified, please label all sites with this problem. As a data user I would not want to include these sites in analysis. Ideally I suggest that the authors exclude investigator-filled data since they are conducting a filling exercise.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we fully agree that investigator-filled data should not be described as "true observational data". We have revised the manuscript and dataset documentation to avoid this wording and to make the distinction between original QA/QC-filtered observations and provider-supplied gap-filled time series explicit.

As noted in the revised manuscript, among the 50 flux tower sites included in this study, 40 sites provide quality-controlled original LE observations, whereas 10 sites only provide complete LE time series that had already been gap-filled by the original data providers within their observation periods. These 10 sites cannot be separated into original observations and investigator-filled values based on the available metadata. Therefore, we have clearly identified these sites in both Appendix Table A1 and the Readme.txt file accompanying the final dataset. This allows users who prefer to use only original QA/QC-filtered observations to easily exclude these sites from their analyses.

We also clarify that all flux tower sites were modeled independently. Therefore, the inclusion of these 10 sites does not affect the model training, gap-filling results, or final data quality of the other 40 sites that provide original QA/QC-filtered observations. For these 10 sites, no additional F-type gap-filled data are generated within the observation period. Only temporally prolonged values beyond the observation period are produced using the AutoML framework and are explicitly flagged as P.

In the final data product, each site is provided as a separate CSV file. Within each file, the QC flag distinguishes different data types: T denotes the LE records provided for the observation period, F denotes LE values gap-filled by our AutoML framework within the observation period, and P denotes LE values produced through temporal prolongation beyond the observation period. For the 40 sites with original observations, T corresponds to original QA/QC-filtered tower observations. For the 10 marked sites, however, T represents provider-supplied complete LE time series within the nominal observation period and should not be interpreted as purely original observations. This distinction is explicitly stated in Appendix Table A1 and in the Readme.txt file.

Although we recognize the reviewer's preference to exclude investigator-filled records, we retained these 10 sites because available flux tower data across China remain limited, and these sites provide additional spatial and ecosystem coverage. Since they are clearly marked and do not influence the results of the other 40 independently modeled sites, retaining them provides users with greater flexibility while allowing users to exclude them when strict use of original observations is required.

We believe this compromise improves transparency and usability while preserving the broader value of the dataset.

The revised version in ***Data availability*** (**the Line 700 to 704** of the revised manuscript, the modified content is displayed in bold):

5 Data availability

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

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

Detailed comments 31

There appear to be five independent gap-filled methods explained. Lines 250-256 explain one approach, based on random removal of data. Section 2.4.2 describes five more approaches, based on elimination of gaps of four different lengths. Each approach is reportedly optimized and applied to each site. How are five different gap filling procedures applied to each flux tower site?

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the previous description could be misunderstood as indicating that multiple independent gap-filling procedures were applied to generate the final dataset for each flux tower site. We have revised the text to clarify the workflow.

In this study, the four artificial gap scenarios, corresponding to gap lengths of 30 min, 1 d, 7 d, and 30 d, were used only for model evaluation and method comparison. They were not four independent production gap-filling procedures. For each site, artificial gaps were introduced into valid observations to test how the model performed under different missing-data durations. This evaluation was repeated 20 times with different random data splits and gap realizations to quantify the robustness and stability of the method.

The benchmark methods, including MDS, RF, and XGBoost, were also applied only under the same artificial-gap evaluation framework. Their purpose was to provide comparative performance statistics under identical input and data-partitioning conditions. They were not used to generate the final released LE time series.

After completing the artificial-gap evaluation, the artificial gaps were discarded. For the final production dataset, each flux tower site was processed independently using one site-specific final AutoML model. This model was retrained using all available high-quality LE observations at that site and was then applied once to fill the actual missing LE values within the observation period. The same final model was subsequently used for temporal prolongation outside the observation period.

Therefore, each flux tower site has only one final production gap-filling model and one final gap-filled LE time series. The multiple artificial-gap scenarios were designed solely to evaluate model performance under different gap-length conditions and do not represent multiple final gap-filling products.

To avoid repetition and ambiguity, we removed the duplicated description from Lines 250–256 and consolidated the full explanation of the training, validation, artificial-gap evaluation, benchmark comparison, and final production gap-filling workflow in Section 2.4.2.

The revised version in ***Data and methodology*** (**the Line 320 to 360** of the revised manuscript, the modified content is displayed in bold):

2.4.2 Artificial gap scenarios

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

using different random data splits and gap realizations to reduce the influence of randomness. **After this evaluation step, the artificial gaps were not used in the final data product. Instead, a single final AutoML model was retrained for each site using all available high-quality LE observations and then applied once to fill the actual missing LE values within the observation period.** This process resulted in seamless half-hourly LE time series, which serve as the baseline dataset for subsequent temporal prolongation and multi-scale analyses.

To facilitate comparative evaluation of different gap-filling approaches, 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. **These benchmark methods were also applied only within the artificial-gap evaluation framework and were not used to generate the final released gap-filled LE dataset. Model performance was assessed using five-fold cross-validation during the training stage to optimize model configuration and reduce the risk of overfitting. After model selection, the optimal model for each site was retrained using all available high-quality LE observations and then applied to fill the missing LE values. Consequently, a site-specific gap-filling model was established for each site, providing a reliable basis for subsequent temporal prolongation of LE time series.**

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

Detailed comments 32

Lines 278-279. Please include citations for these "benchmark methods." If these are routinely used for flux tower gap filling, please provide the citations documenting this work. Please provide citations for the methods you are using.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we have revised the manuscript to include references documenting the use of each method in the context of eddy covariance flux gap filling.

Specifically, we cite Foltýnová et al. (2020) to document the marginal distribution sampling (MDS) approach, which has been widely recommended for half-hourly latent heat flux gap filling. We also cite Khan et al. (2021) for the application of random forest (RF) regression in gap-filling latent heat flux, and Liu et al. (2025) for the use of extreme gradient boosting (XGBoost) to robustly fill long gaps in eddy covariance flux measurements. These citations provide detailed examples of how each method has been applied in the literature and improve the documentation of the methods used as benchmarks in our comparative evaluation.

The revised version in ***Data and methodology*** (**the Line 334 to 336** 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 **(Foltýnová et al., 2020)** and two commonly used machine learning algorithms, random forest (RF) **(Khan et al., 2021)** and extreme gradient boosting (XGBoost) (**Liu et al., 2025**), were implemented as benchmark methods.

*The **Suplementary References***:

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

Khan, M. S., Jeon, S. B., & Jeong, M.-H.: Gap-Filling Eddy Covariance Latent Heat Flux: Inter-Comparison of Four Machine Learning Model Predictions and Uncertainties in Forest Ecosystem, Remote Sensing, 13(24), 4976, https://doi.org/10.3390/rs13244976, 2021.

Liu, Y., Lucas, B., Bergl, D. D., & Richardson, A. D.: Robust filling of extra-long gaps in eddy covariance CO_2 flux measurements using eXtreme Gradient Boosting, Agricultural and Forest Meteorology, 364, 110438, https://doi.org/10.1016/j.agrformet.2025.110438, 2025.

Detailed comments 33

Line 290. Three metrics are presented, with citations, this is helpful. The way that these are used to determine optimal performance, however, is not explained. Please explain. (The text notes that four metrics are employed, but only three are mentioned in the text. Please correct this error.)

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the original manuscript did not clearly explain how model optimality was determined and also contained an inconsistency regarding the number of evaluation metrics.

First, we corrected the text to indicate that three, rather than four, performance metrics were used: the correlation coefficient (CC), root mean square error (RMSE), and percentage bias (PBias).

Second, we have clarified the role of each metric in the model evaluation and selection process. In this study, the optimal model was determined solely based on RMSE calculated on the validation dataset. This choice is consistent with the default ranking strategy of the H2O AutoML framework for regression problems and is widely adopted in flux gap-filling studies. RMSE directly measures the magnitude of prediction errors in the same physical units as latent heat flux (W m^{-2}), providing an intuitive and robust indicator of overall predictive accuracy. Because RMSE penalizes large errors more strongly than small errors, it is particularly suitable for identifying models that best reproduce observed LE variability.

By contrast, CC and PBias were not used for model selection. Instead, they were reported as complementary performance metrics. CC was used to evaluate the strength of the relationship between predicted and observed values, whereas PBias was used to assess systematic overestimation or underestimation. Together, these metrics provide a more comprehensive characterization of model performance beyond overall prediction error.

To further improve transparency and reproducibility, we note that the complete workflow implementation, including the model selection procedure and evaluation metrics, has been publicly released and is available through our open-access code repository (https://doi.org/10.5281/zenodo.18194590; Qian et al., 2026).

The manuscript has been revised accordingly.

The revised version in ***Data and methodology*** (**the Line 359 to 362** of the revised manuscript, the modified content is displayed in bold):

In addition, **three** commonly used performance metrics were employed to quantify gap-filling performance, including the correlation coefficient (CC), root mean square error (RMSE, W m-2), and percentage bias (PBias, %). The definitions of CC and RMSE follow Li et al. (2025a), while PBias was calculated according to Qian et al. (2023). **And RMSE was used as the model selection criterion, whereas CC and PBias were reported for supplementary performance evaluation.**

Detailed comments 34

Section 2.4.3. Similar to the previous section, a number of strategies for testing the prolongation results are described but the performance metrics are not specified. In addition, the decisions made based on the comparison of performance metrics are not specified. Testing the methodology is good, but explaining the metrics and explaining how the tests will be used is also necessary.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the original text did not explicitly describe which performance metrics were used to evaluate temporal prolongation results and how these metrics informed model decisions.

We have revised Section 2.4.3 to clarify that the testing of prolongation models uses three metrics: correlation coefficient (CC), root mean square error (RMSE), and percentage bias (PBias), consistent with the metrics used in Sect. 2.4.2 for gap-filling evaluation. Among these, RMSE was used as the primary criterion to identify the optimal prolongation model for each site, as it directly reflects prediction error in the same physical units as latent heat flux. CC and PBias were reported as complementary metrics to evaluate correlation and systematic bias. This approach allows us to objectively assess both directional consistency (backward vs. forward prolongation) and temporal stability under different training-length scenarios, while providing transparent and reproducible evaluation criteria.

In addition, all prolongation and evaluation procedures were implemented per site, independently, using the publicly released workflow code (https://doi.org/10.5281/zenodo.18194590; Qian et al., 2026), ensuring full reproducibility and traceability. The manuscript now explicitly states the metrics used and how they were applied in model selection and evaluation.

We believe these revisions address the reviewer's concern by specifying both the performance metrics and their role in determining optimal models for temporal prolongation.

The revised version in ***Data and methodology*** (**the Line 363 to 389** of the revised manuscript, the modified content is displayed in bold):

2.4.3 Prolonged hourly LE data

After completing seamless gap-filling within the observation period, the latent heat flux (LE) time series at each flux tower site were further temporally prolonged at the half-hourly scale to construct continuous records covering a unified time span. Temporal prolongation was also implemented within the AutoML of H2O framework, using the same model architecture, input feature set, and training strategy as described in Sect. 2.4.1 to ensure methodological consistency. For each site, a site-specific prolongation model was independently trained using the available LE observations as supervisory information. Five-fold cross-validation was applied to identify the optimal model configuration and hyperparameter combination. Temporal prolongation was performed in two directions at the half-hourly scale. For a site with an observation period spanning a specific interval (e.g., 2006–2015), the period preceding the observations (e.g., 2000–2005) was defined as backward prolongation,

whereas the period following the observations (e.g., 2016–2022) was defined as forward prolongation. Under identical data and modeling conditions, the prolongation performance in both temporal directions is expected to be comparable. To evaluate the consistency of prolongation results across temporal directions, a symmetric data partitioning strategy was adopted. For backward prolongation, the last two-thirds of the observed time series were used for model training, while the first one-third served as the testing set. Conversely, for forward prolongation, the first two-thirds of the data were used for training and the remaining one-third for testing. **Model performance on the testing sets was evaluated using CC, RMSE, and PBias. Optimal prolongation models were selected based on the lowest RMSE, while CC and PBias served as complementary metrics to assess correlation and systematic bias. This approach enables an objective evaluation of directional consistency and the detection of potential systematic deviations during temporal prolongation.**

In addition, to examine the temporal stability of model performance as the length of the prolongation period increases, two representative training-length scenarios were designed. For sites with observation periods of at least 6 years, the first 6 years of data were used for training and the remaining years for testing. For sites with observation periods of at least 3 years, the first 3 years were used for training, with subsequent years reserved for testing. By comparing prolongation performance under different training-length conditions, **using the same metrics (CC, RMSE, PBias) and selecting models based on RMSE,** the long-term stability and reliability of the model during temporal extension were systematically evaluated. All prolongation and validation procedures were conducted independently at the site scale. After completing consistency and stability assessments, the final model for each site was retrained using all available high-quality LE observations and applied to periods outside the observation interval, generating continuous half-hourly LE time series covering 2000–2024.

Detailed comments 35

Figure 3. The authors present box and whisker plots but have not specified the population being analyzed. Is there a single value of each metric for each of the flux towers? This must be explained in the figure caption.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions. The figure presents the performance metrics (CC, RMSE, PBias) for each flux tower as the mean values across 20 repeated experiments under each artificial gap scenario. Therefore, each tower contributes a single averaged value per metric in the figure. The use of repeated experiments with different random data splits and gap realizations allows for robust assessment of model performance, while the figure summarizes the central tendency of these results for ease of visualization and comparison among methods.

To further clarify the uncertainty associated with the AutoML-based gap-filling, Section 4.2 was added to discuss the distributions of CC, RMSE, and PBias across the 20 repeated experiments. This analysis demonstrates that the dispersion of all three metrics remains limited, with higher uncertainty observed for longer continuous gaps (e.g., 30-day gaps) and lower uncertainty for short gaps (e.g., 30 min). These results provide empirical uncertainty bounds, informing users of the reliability of the gap-filled data and highlighting the statistical stability of the proposed AutoML framework.

The figure caption has been updated to explicitly explain the population analyzed, and Section 4.2 provides a detailed discussion of the repeat-experiment ranges and associated uncertainties, ensuring full transparency and reproducibility of the performance assessment.

The revised ***title of Figure 3***:

Figure 3: Comparison of half-hourly LE gap-filling performance among different methods under various artificial gap scenarios. **For each flux tower, each metric (CC, RMSE, PBias) represents the mean value across 20 repeated experiments. Each site thus contributes a single averaged value for each metric in the figure.**

The revised version in ***Discussion*** (**the Line 628 to 651** of the revised manuscript, the modified content is displayed in bold):

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.

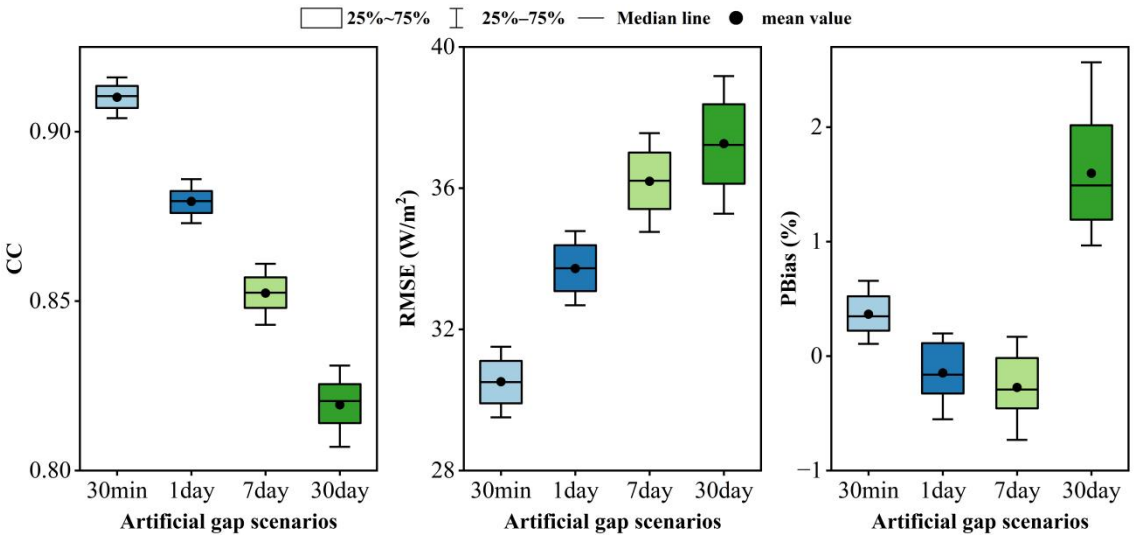

Figure 12: Uncertainty of AutoML-based half-hourly LE gap-filling performance under different artificial gap scenarios (30 min, 1 day, 7 day, and 30 day), estimated from 20 repeated experiments with different random data splits and gap realizations. Boxplots show the distributions of correlation coefficient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

Detailed comments 36
Figure 4 presents five different versions of AutoML gap filling. Which one is used to construct the data set presented by this manuscript? Or does this manuscript present a data set that uses five different gap-filling methods? If so, what guidance is presented for potential users of these data?

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the original description of Figure 4 could cause confusion regarding whether multiple AutoML gap-filling products were generated.

We clarify that the dataset presented in this manuscript was constructed using only one final site-specific AutoML model for each flux tower site. The five AutoML-related entries shown in Figure 4 do not represent five different released gap-filling datasets. Instead, they correspond to performance evaluation results under different artificial gap scenarios, including 30 min, 1 d, 7 d, 30 d, and all-gap scenarios. These scenarios were designed only to assess the robustness of the AutoML framework under different missing-data lengths and environmental conditions.

After the artificial-gap evaluation was completed, the artificial gaps and scenario-specific evaluation models were not used to generate the final released dataset. For the final data production, a single optimal AutoML model was retrained for each site using all available high-quality LE observations. This final site-specific model was then applied once to fill actual missing LE values within the observation period and to conduct temporal prolongation outside the observation period.

Therefore, the released dataset does not contain five different gap-filled versions. Each site has one final continuous LE product, with the QC flag indicating whether each value is observed, gap-filled within the observation period, or temporally prolonged. To avoid ambiguity, the caption of Figure 4 and the related text have been revised to explicitly state that Figure 4 presents only evaluation results under different artificial gap scenarios and does not represent multiple final data products.

We also clarified this point in the manuscript to guide potential users. Users should use the final released LE time series for each site as the recommended dataset, while Figures 3 and 4 provide methodological performance evaluation and uncertainty information supporting the reliability of the final product.

The revised ***title of Figure 4***:

Figure 4: Performance of different methods across underlying surface types and climate zones under varying artificial gap scenarios. **The results shown here represent evaluation outcomes under different artificial gap-length scenarios and do not correspond to separate released gap-filled datasets. The final released dataset was generated using one site-specific optimal model for each flux tower site.**

Detailed comments 37

Results. The results immediately focus on comparing AutoML to other methods without showing any of the results of the process used (e.g. sections 2.4 and 2.5 of the methods) to optimize the AutoML approach.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the original Results section moved too quickly to the comparison between AutoML and benchmark methods, without first presenting sufficient results from the AutoML optimization and evaluation procedures described in Sections 2.4 and 2.5. This could make it unclear how the final AutoML framework was selected and why it was subsequently used to generate the released dataset.

To address this concern, the Results section has been expanded to better reflect the evaluation procedures described in the Methods section before presenting comparisons with benchmark methods. Specifically, a new subsection (Section 4.2) was added to report the uncertainty assessment of the AutoML framework based on 20 repeated experiments under different artificial gap scenarios. The distributions of CC, RMSE, and PBias are presented and discussed, providing direct evidence of the robustness and stability of the AutoML approach under different gap-length conditions. These additional results establish the reliability of the AutoML framework prior to the comparison with benchmark methods and provide a direct link between the evaluation procedures described in Sections 2.4 and 2.5 and the final dataset generation process.

Detailed comments 38

Lines 374-375. "This pattern reflects the increased uncertainty of evapotranspiration processes under humid and subtropical climate conditions."

Your RMSE metric is dimensional. Performance based on these metrics will also be a function of the magnitude of the flux. It is not surprising that the RMSE for arid sites is smaller than the same metrics for humid subropical sites.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that interpreting regional differences solely based on RMSE may be incomplete because different performance metrics characterize different aspects of model behavior. In response, the revised manuscript now considers CC, RMSE, and PBias together when describing regional performance patterns.

Specifically, the revised text highlights that the Inner Mongolia Temperate Semiarid Zone (IMSZ) and the Northeast Temperate Subhumid Zone (NETSZ) generally exhibit higher CC and lower RMSE values, indicating stronger predictive skill. In contrast, the Southeast Subtropical Humid Zone (SESHZ), Southern Tropical Humid Zone (STHZ), and Northern Temperate Subhumid Warm Zone (NTSWZ) tend to show lower correlations and larger reconstruction errors. At the same time, the results also demonstrate that regions with favorable CC and RMSE do not necessarily exhibit the smallest systematic bias. For example, although the Northwest Desert Arid Zone (NWDAZ) shows relatively strong correlation performance, its PBias is larger than that of several other climate zones.

Accordingly, the revised interpretation avoids relying on a single metric and instead emphasizes that the strengths and limitations of the gap-filling model vary depending on whether performance is evaluated in terms of correlation (CC), absolute error (RMSE), or systematic bias (PBias). This provides a more balanced and comprehensive assessment of model performance across different climatic regions.

The revised version in ***Discussion*** (**the Line 447 to 455** of the revised manuscript, the modified content is displayed in bold):

In terms of regional differences, gap-filling performance is relatively better in the Inner Mongolia Temperate Semiarid Zone (IMSZ) and the Northeast Temperate Subhumid Zone (NETSZ), characterized by higher correlations and lower errors. **For example, AutoML achieves an overall CC of 0.839, RMSE of 22.69 W m^{-2}, and PBias of 0.50 % in IMSZ, while NETSZ exhibits the highest overall CC (0.910) with a relatively low RMSE (28.28 W m^{-2}). In contrast, the Southeast Subtropical Humid Zone (SESHZ), Southern Tropical Humid Zone (STHZ), and Northern Temperate Subhumid Warm Zone (NTSWZ) generally show lower CC values and higher RMSE values, indicating reduced reconstruction accuracy under these climatic conditions. However, although arid and semiarid regions generally exhibit stronger correlations and lower RMSE values, some regions, such as the Northwest Desert Arid Zone (NWDAZ), show comparatively larger systematic deviations, with PBias reaching. Overall, the results indicate that model performance varies across climate zones, with different regions exhibiting distinct strengths and limitations when assessed from the perspectives of correlation, absolute error, and systematic bias.**

Detailed comments 39

Section 3.1.2 introduces a metric, the ability to reproduce the shape of the diel cycle, that is not defined in the methods. This metric needs to be explained in the methods. The reason for selecting this metric should also be described in the methods. Why is this metric added?

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we clarify that this is not a formal statistical metric, but rather a qualitative, visual assessment of temporal fidelity. Its purpose is to evaluate whether the

gap-filling methods preserve the natural diurnal patterns and amplitude variations of latent heat flux (LE) at the half-hourly scale, especially during extended missing periods. This assessment helps demonstrate that the reconstructed LE time series remain physically plausible and retain the characteristic temporal structures observed in the original data, which cannot be fully captured by pointwise metrics such as RMSE, CC, or PBias alone.

The use of this qualitative evaluation of diel cycle reconstruction follows precedent in flux gap-filling studies, including ESSD 2025 (Li et al., 2025, https://doi.org/10.5194/essd-17-3835-2025), where visual inspection of diurnal patterns across different surface types and climates is used to supplement statistical metrics for assessing model performance. This approach provides insight into structural distortions, smoothing effects, or amplification of noise that might occur under long continuous gap conditions.

In the revised manuscript, we have clarified in the Methods section that this evaluation represents a visual/structural assessment, explained its rationale, and linked it to the referenced ESSD study. Users should interpret this assessment as a complementary check for temporal consistency and physical plausibility rather than a numerical metric. All other quantitative performance assessments, including RMSE, CC, and PBias, remain the primary criteria for model comparison and selection.

This clarification ensures that readers understand both the purpose and the limitations of the diel cycle assessment, and prevents misinterpretation of it as a formal statistical metric.

The revised version in ***Data and methodology*** (**the Line 361 to 369** of the revised manuscript, the modified content is displayed in bold):

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

Detailed comments 40-42

40. Figure 5. The gaps are not 30 days long. Please explain how this is an illustration of 30-day gap filling.

41. Figure 5. The comparison across methods is not clear. I suggest a more statistical evaluation of the daily cycle - a mean daily cycle and variability, for example. I cannot support the conclusions in the text regarding the superiority of the AutoML results based on the results shown in this format.

42. Figure 5. It appears (e.g. Fig. 5a) that AutoML sometimes underestimates the observed fluxes and creates a very smoothed representation of the fluxes. It would be helpful if the authors could describe whether or not this method creates a time series that is artificially smoothed in comparison to true flux measurements. True flux measurements include random sampling error (e.g. Lenschow and Stankov, 1986; Richardson et al, 2006). If AutoML does not reproduce this inherent feature of an EC flux time series this should be described.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we provide the following clarification and additional context to address all three concerns in a unified manner.

The time intervals displayed for each site in Figures 5 and 6 correspond precisely to the artificially imposed 30-day continuous gaps used for validation. During model development, all observations within these periods were completely withheld and were not available to either the AutoML or MDS methods. The observed LE values shown within these intervals were added back solely for visual comparison and validation purposes. In particular, the blue dashed boxes highlight periods where MDS exhibited pronounced deviations, whereas AutoML maintained consistent reconstruction, emphasizing the stability of AutoML under challenging missing-data conditions.

To provide a more quantitative evaluation of reconstruction quality beyond visual inspection, two descriptive statistics were calculated for each representative site: the mean LE and the coefficient of variation (CV, %) over the 30-day gap period. These statistics complement the previously described diel-cycle assessment, allowing for a systematic comparison of methods in terms of central tendency and temporal variability. Across the majority of sites and underlying surface types, AutoML better reproduces both the observed mean LE and variability compared to MDS, while retaining physically plausible diurnal cycles and amplitude patterns.

Regarding the smoothness of the AutoML reconstruction, we note that no explicit temporal smoothing or post-processing was applied during model training or prediction. The smoother appearance of some AutoML reconstructions arises naturally from the machine-learning model capturing dominant environmental controls and filtering high-frequency stochastic noise inherent in eddy covariance measurements. Consequently, although minor high-frequency fluctuations may be attenuated, the primary temporal structure, diurnal cycles, and physical plausibility of the LE series are preserved.

Overall, this unified assessment demonstrates that under long continuous gaps, AutoML consistently maintains temporal continuity, reproduces diurnal and monthly patterns, and preserves both mean state and variability across climate zones and land-cover types. The revised figures, figure captions, and text now explicitly clarify the 30-day gap representation, the role of blue dashed boxes, and the complementary statistical evaluation, thereby addressing the reviewers' concerns regarding gap length, comparison clarity, and smoothness.

The revised version in ***Data and methodology*** (**the Line 369 to 373** of the revised manuscript, the modified content is displayed in bold):

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

The revised version in ***Result*** (**the Line 471 to 501** of the revised manuscript, the modified content is displayed in bold):

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

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

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

Detailed comments 43

Figure 7. The population of points is not described. Please explain what each point on (a) or (b) represents, and how many points make up (a) through (f).

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, the population represented in Figure 7 originates from the temporal prolongation evaluation procedure described in Section 2.4.3.

Specifically, for each flux tower site, a symmetric data partitioning strategy was adopted to evaluate directional consistency during temporal prolongation. For backward prolongation, the last two-thirds of the observed LE time series were used for model training and the first one-third for testing. For forward prolongation, the first two-thirds were used for training and the remaining one-third for testing. The prolonged LE values generated for the testing periods were then compared against the corresponding observed LE values.

Accordingly, each point in Figures 7a and 7b represents one half-hourly LE record from the testing dataset, plotted as prolonged LE versus observed LE. The scatter-density plots therefore summarize all half-hourly testing observations from all sites included in the backward and forward prolongation experiments, respectively. To improve clarity, the total number of data points (N) has now been added directly to Figures 7a and 7b.

We also note that Figures 7c–f do not represent individual data points. Instead, these panels summarize the mean performance of the prolongation models across different underlying surface types and climate zones. Panels 7c and 7e show the mean CC and RMSE values for backward and forward prolongation across land-cover categories, while Panels 7d and 7f present the

corresponding mean values across climate zones. Therefore, these panels are aggregated performance summaries rather than scatter plots of individual observations.

To avoid ambiguity, the figure caption has been revised accordingly.

The revised ***Figure 7***:

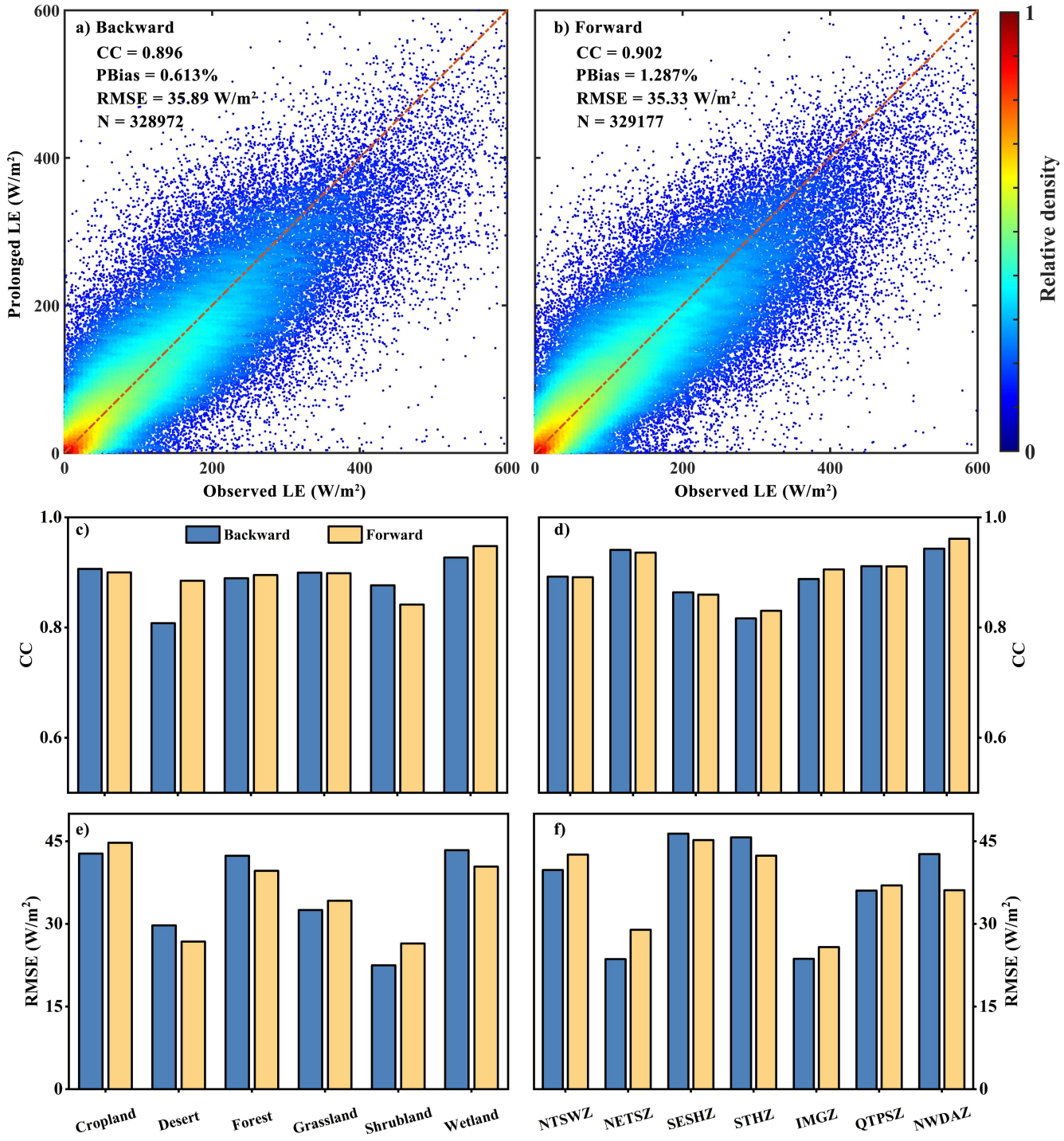

Figure 7: Consistency between backward (a) and forward (b) prolongation of half-hourly latent heat flux (LE), shown as scatter density plots of prolonged LE versus observed LE. **Each point represents one half-hourly LE record from the testing dataset used in the prolongation validation, and N denotes the total number of testing records included in each panel.** The color scale

represents the relative density of data points, normalized to highlight areas with higher concentration (warmer colors) and lower concentration (cooler colors). Panels (c) and (e) show the correlation coefficient (CC) and root mean square error (RMSE) across different underlying surface types, respectively, while panels (d) and (f) present the corresponding metrics across climate zones.

Detailed comments 44

Figure 8. Please define the x-axis more precisely in the figure caption and/or axis label. "Year" is not clear enough to be readily and rapidly understood.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions. The x-axis in Figure 8 represents the calendar year of the testing period used to evaluate forward prolongation performance. To clarify, for sites with at least 6 years of observations, the first 6 years were used to train the model and the subsequent years (7th year onward) were used as testing data. For sites with at least 2 years, the first 2 years were used for training and the remaining years were used for testing. Each bar (RMSE) and line (CC) in panel (a) corresponds to the performance of the site-specific AutoML model on the testing data of that particular year. This design allows for a year-by-year assessment of temporal stability in the model's predictions as the prolongation period increases. The figure caption has been updated to explicitly convey this meaning.

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). The x-axis "Year" indicates the calendar year of testing data outside the training interval: for sites with six or more years of observations, the first six years were used for training and subsequent years (seventh year onward) for testing; for sites with two or more years, the first two years were used for training and subsequent years for testing. Each bar and line represents the performance of the model on that year of testing data.**

Detailed comments 45-52

45. Lines 454-455. "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."

46. These claims are very qualitative. The simple time series plots in Figure 9 are not sufficient to justify quantitative statements about the performance of this data product. I suggest that more quantitative metrics be used to illustrate the quality of the reproduction of the diel cycle of fluxes.

47. Lines 465-467. "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."

48. Section 3.3.2. A time series comparison is an interesting visual but it does not provide quantitative understanding regarding the performance of the AutoML model. I am uncertain of the value of showing a subset of sites and only a qualitative comparison.

49. Figure 10. Some of the sites show relatively poor agreement with the model either in particular years ((g), 2014) or over the entire sequence (daily correlation, (f)). This isn't discussed in the text.

50. Lines 492-493. "Fig. 11 presents time series comparisons between prolonged monthly LE and observation-based aggregated LE at several representative sites."

Text of this variety belongs in the figure caption. Please remove all descriptions of the figures from the text and place them in the figure caption. This issue exists at many points in the document.

51. Lines 500-501. "some uncertainty remains in monthly LE estimates."

This is not a helpful statement. Please explain "some uncertainty."

52. Figure 11. I am uncertain of the value of this figure in the manuscript.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the original version of Section 3.3 relied heavily on visual inspection of the representative time-series examples shown in Figure 9 and that the supporting quantitative evidence was insufficient.

To address this concern, we revised Figure 9 by adding three quantitative performance metrics, namely the correlation coefficient (CC), root mean square error (RMSE), and percentage bias (PBias), for each representative site. These metrics provide an objective evaluation of the ability of the prolongation model to reproduce observed LE dynamics, including temporal variability and diel-cycle characteristics, and complement the visual comparison of the half-hourly time series. The revised figure therefore combines both graphical and statistical information when assessing prolongation performance.

In addition, to improve the overall organization and readability of the manuscript, and in response to recommendations from Reviewer #1 regarding redundancy and manuscript length, Figures 10 and 11 have been moved to the Appendix. Correspondingly, the text in Section 3.3 has been reorganized and streamlined. Rather than separating the results into multiple subsections, the revised manuscript now presents the prolongation evaluation within a single integrated Section 3.3. This revision improves readability while retaining all essential methodological and validation information.

We also agree that the original statement regarding the reproduction of "high-frequency LE variability" was stronger than warranted by the evidence presented in Figure 9 alone. Accordingly, the text has been revised to avoid over-interpreting the representative examples. The role of Figure 9 is now described as providing representative evidence that the prolongation framework can reasonably reproduce the major temporal patterns and diel-cycle characteristics of observed LE, while the newly added CC, RMSE, and PBias metrics provide quantitative support for the quality of the reconstruction. This revision ensures that the conclusions are supported by both visual and statistical evidence and avoids relying solely on qualitative interpretation.

In addition, we acknowledge that the original Section 3.3.2 and Figures 10 and 11 relied heavily on visual inspection of a subset of sites, which may have limited the quantitative understanding of the AutoML model's performance. To address this concern, Figures 10 and 11 have been moved to the Appendix, and the main text of Section 3.3 has been reorganized into a single integrated subsection to improve readability and reduce redundancy.

We emphasize that while some sites in Figures 10 and 11 show relatively lower agreement in particular years or over the entire sequence (e.g., daily correlation in panel f or the 2014 data in panel g), these variations are acknowledged in the revised text and are consistent with expected variability due to differences in observational coverage, site-specific microclimatic conditions, and longer-term temporal dynamics. The discussion now highlights that the quantitative metrics CC, RMSE, and PBias—added to Figure 9—provide objective support and complement the qualitative visual comparisons, ensuring that performance evaluation is not based solely on representative examples.

Finally, to improve clarity and adhere to journal guidelines, all detailed descriptions that previously appeared in the main text have been moved into the respective figure captions. This ensures that the text focuses on summarizing overall results and interpretations, while figure captions contain figure-specific explanations.

Overall, these revisions collectively enhance the quantitative and visual support for the prolongation framework, address concerns regarding over-reliance on visual inspection, acknowledge sites with lower agreement, and improve the clarity and organization of the manuscript.

The revised ***Section 3.3*** and ***Figure 9***:

3.3 Demonstration of different scale prolonged time series

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

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

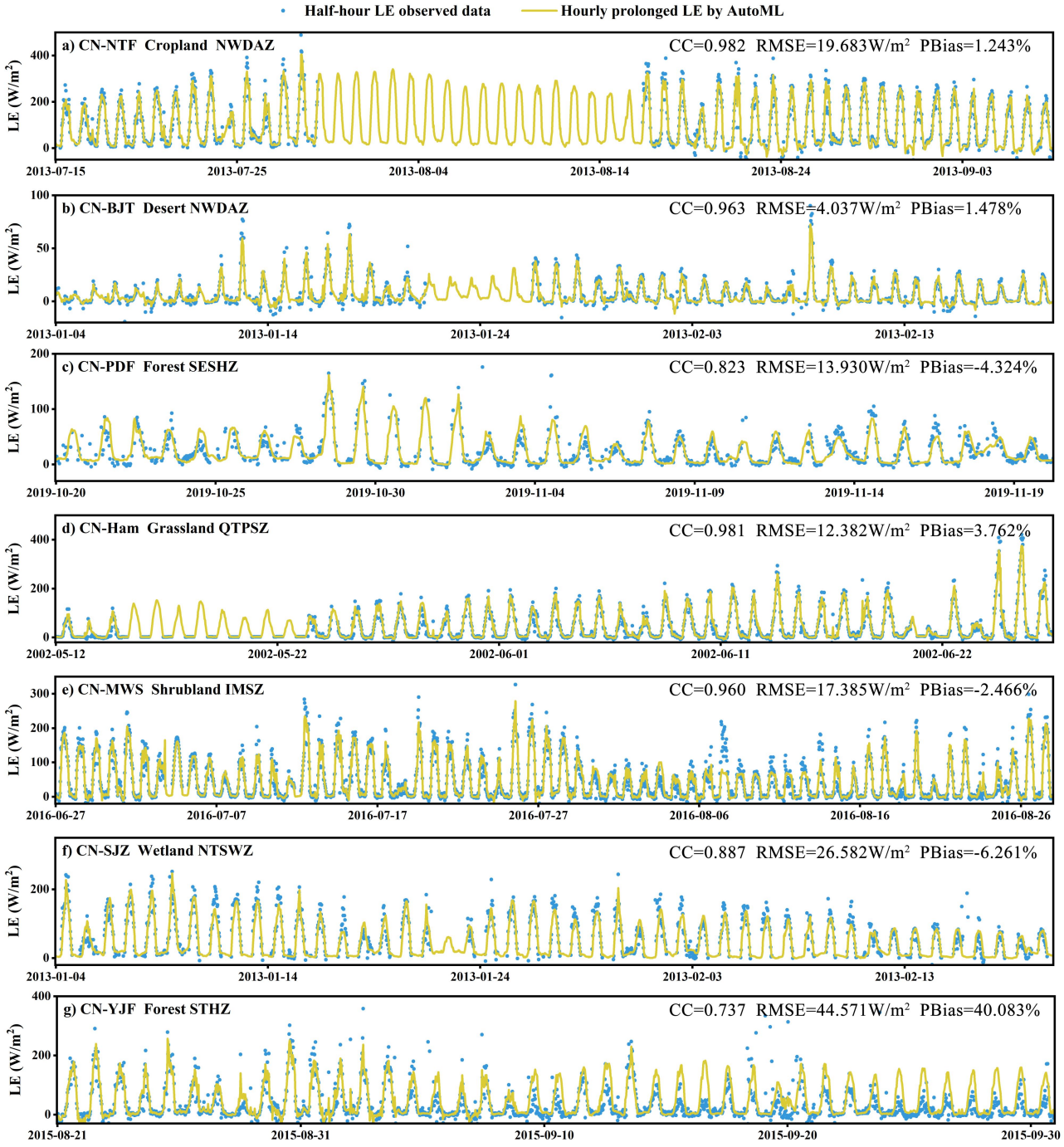

Figure 9: Demonstration of half-hourly prolonged LE time series across different typical sites.

Detailed comments 53

Lines 518-522. "The medians and distribution ranges of the two datasets are comparable, indicating that the gap-filling and temporal prolongation procedures do not introduce evident systematic biases. This consistency is generally stable across most underlying surface types and climate zones, with relatively larger dispersion observed only in high-variability ecosystems such as desert and shrubland, yet without any directional bias."

"Comparable", "high-variability ecosystems" and "relatively larger dispersion" are all vague terms. "high consistency" is also used earlier in this paragraph. These are not useful analyses. The degree to which this data product represents true observations should be described concisely and quantitatively.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the original text relied on qualitative expressions such as "high consistency", "comparable", and "relatively larger dispersion", which do not provide a sufficiently objective description of the agreement between the reconstructed dataset and the ChinaFlux observations.

Following the reviewer's suggestion, we revised this paragraph to describe the comparison in terms of specific distributional characteristics directly supported by Fig. 12. In the revised manuscript, we explicitly refer to the overlap of median values, interquartile ranges (25th–75th percentiles), and overall distribution spreads between the two datasets. We also replaced the subjective expression "high-variability ecosystems" with a direct description of the observed statistical behavior of desert and shrubland sites, namely their wider distribution ranges and larger interquartile spreads.

Because Fig. 10 is a boxplot-based comparison rather than a point-to-point accuracy assessment, we avoided introducing unsupported quantitative accuracy metrics (e.g., CC or RMSE) in this section. Instead, the revised text focuses on the distributional agreement that can be directly inferred from the figure. The updated paragraph now reads:

The revised version in ***Discussion*** **(the Line 591 to 602** of the revised manuscript, the modified content is displayed in bold):

4.1 Comparison between ChinaFlux and our dataset

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

Detailed comments 54-55

54. Figure 13. I am concerned about the degree to which the "ChinaFlux" data (the x-axis) might be a filled data product. This is especially true at monthly to annual time scales where gap filling is essential. These become comparisons between gap filling algorithms, not between observations and a gap-filling algorithm. At minimum I would require any data on the x-axis to have a minimum fraction of true observations, and for that threshold to be stated clearly in the text. I would also eliminate all gap-filled data from the hourly data product comparison.

55. Lines 526-527. Please move this text to the figure caption.

Author Respond: We thank the reviewer for this important comment and apologize for not describing the screening criteria for the ChinaFlux reference data sufficiently clearly in the original manuscript.

In fact, quality-control thresholds had already been applied prior to the daily, monthly, and yearly comparisons, but these criteria were only described in Sections 3.3.2 and 3.3.3 and were not explicitly referenced in the discussion associated with Fig. 11. Specifically:

For the daily-scale comparison, only days with less than 10 % missing half-hourly LE observations were retained as valid samples.

For the monthly-scale comparison, only months with less than 10 % missing daily LE observations were retained.

For the yearly-scale comparison, only years with less than 10 % missing daily LE observations were retained.

Therefore, the ChinaFlux data used on the x-axis of Fig. 11 were not generated from heavily gap-filled records but were restricted to periods dominated by direct observations. The purpose of these screening criteria was precisely to minimize the influence of gap-filled values and ensure that the comparison remained as close as possible to an observation-based evaluation.

To make this clearer, we have added the above information directly to the Figure 11 caption, where readers can immediately identify the data-screening criteria used for the comparison.

In addition, following the reviewer's suggestion, the figure-descriptive text previously included in the main body (Lines 526–527) has been removed from the manuscript and incorporated into the figure caption. We further reviewed the manuscript and removed or relocated similar figure-descriptive statements where appropriate, which has improved the overall readability of the paper.

The revised ***title of Figure 11***:

Figure 11: Comparison between the reconstructed LE dataset and ChinaFlux observations aggregated at daily (a,d), monthly (b,e), and yearly (c,f) scales for different underlying surface types and climate zones. **For daily-scale comparisons, only days with less than 10 % missing half-hourly LE observations were retained as valid samples. For monthly-scale comparisons, only months with less than 10 % missing daily LE observations were included. For yearly-scale comparisons, only years with less than 10 % missing daily LE observations were considered valid. These screening criteria were applied to ensure that the ChinaFlux reference data were primarily based on direct observations rather than gap-filled values.**

Detailed comments 56

56. Lines 527-529. "As the temporal scale progresses from daily to monthly and annual, the agreement between the two datasets further improves and dispersion decreases markedly, suggesting that the prolongation results effectively preserve the evapotranspiration characteristics reflected by ChinaFlux observations in a long-term mean sense."

The reduction in variability with time averaging is true when averaging flux observations (and many other data) and is not a metric that the algorithm is effective.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the reduction in variability from daily to monthly and annual timescales is an expected consequence of temporal aggregation and should not, by itself, be interpreted as evidence of algorithm performance. The original wording could inadvertently imply such a causal relationship.

To address this concern, we revised the text to distinguish between the statistical effects of temporal aggregation and the actual evaluation of the prolongation framework. In the revised manuscript, the reduction in dispersion is described as an expected outcome of temporal averaging rather than a performance indicator. Instead, the discussion now focuses on the persistence of agreement between the reconstructed dataset and ChinaFlux observations across aggregation scales, including the consistency of median values, distribution ranges, variation magnitude, and interannual trends.

We further clarify that the key advantage of the reconstructed dataset is not the reduction of variability itself, but its ability to provide temporally continuous estimates during periods affected by observational gaps while maintaining consistency with the

available observations. The monthly and annual comparisons demonstrate that the prolongation procedure preserves the large-scale temporal characteristics represented in the original measurements without introducing substantial systematic deviations.

The manuscript has been revised accordingly to avoid over-interpreting the effects of temporal aggregation and to provide a more rigorous interpretation of the comparison results.

The revised version in ***Discussion*** **(the Line 608 to 619** of the revised manuscript, the modified content is displayed in bold):

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

Detailed comments 57

Lines 536-540. This is a generic statement that is not helpful. Please delete.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the original statement in Lines 536–540 was generic and did not provide additional insight. In response, this sentence has been removed from the revised manuscript to improve clarity and conciseness. The surrounding text has been slightly adjusted to ensure smooth continuity without loss of content.

Detailed comments 58-59

Figure 14. The quantitative metrics in this figure are not defined. the population of points (?) shown within the figure is not defined. More information is needed for this to be useful in the manuscript.

Figure 15. Same comments as for figure 14.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the original figure captions did not provide sufficient information for readers unfamiliar with SHAP analysis to fully interpret the quantitative metrics and graphical elements presented in Figures 13 and 14.

To improve clarity, the figure captions have been revised to explicitly define all quantitative metrics and graphical components. Specifically, we now explain that the bar charts represent the mean absolute SHAP values (Mean(|SHAP value|)), which quantify the average contribution of each predictor to model output. The percentages shown beside each feature indicate its relative contribution to the total SHAP importance within the corresponding model.

We also clarify the meaning of the points displayed in the beeswarm plots. Each point represents one half-hourly sample used in model interpretation. The horizontal position of each point corresponds to its SHAP value, indicating the magnitude and direction of the predictor's contribution to the LE prediction for that sample. Positive SHAP values increase the predicted LE, whereas negative SHAP values decrease it. The color scale represents the relative magnitude of the predictor value, ranging from low values (purple) to high values (yellow).

These revisions provide the necessary methodological context for interpreting both feature importance and feature contribution patterns and improve the accessibility of the figures for readers from different disciplinary backgrounds.

The revised ***title of Figure 13 and Figure 14***:

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

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

Detailed comments 60

Section 4.3. This section says nothing that I find valuable to publish. I don't object with the statements, but they bring no new or insightful information to the reader.

Author Respond: Thank you for pointing out the issues and providing valuable suggestions, we agree that the original Section 4.3 ("Contributions, Limitations, and Future Prospects") was largely descriptive and did not provide substantial new insights beyond information already presented elsewhere in the manuscript.

Following the reviewer's suggestion, we have removed the entire Section 4.3 from the revised manuscript. The key methodological limitations and dataset applicability have already been discussed in the relevant sections of the paper, and therefore the standalone discussion section was not essential for understanding or using the dataset.

We believe that removing this section improves the overall conciseness and readability of the manuscript by avoiding repetition and allowing the manuscript to focus more directly on the dataset construction, validation, and data records.

Detailed comments 61-62

61. Section 5. Auxiliary data. I support the inclusion of the ERA-5 and MODIS data at the sites that was used to create the LE data product. I would strongly suggest, however, that the site-level observations (e.g. of radiative fluxes, sensible heat flux, atmospheric and soil state variables) also be included in this data product.

62. I do not see any indication that site metadata is provided in the data set. A description of the site ecosystem, terrain and soil characteristics is very important for data interpretation. Please add these metadata to the data product.

Author Respond: We thank the reviewer for this valuable suggestion and fully agree that site-level observations and metadata are important for the interpretation and potential reuse of the dataset.

In fact, these supporting resources have already been incorporated into the data documentation through Appendix Table A2, although this was not sufficiently emphasized in the original manuscript. Specifically, Table A2 provides both (1) data-access links and (2) reference publication links for all ChinaFlux sites included in this study.

Through the data-access links, users can obtain the original site-level observations maintained by the corresponding site teams and data repositories. Depending on site availability, these records include not only latent heat flux observations but also additional measurements such as radiative fluxes, sensible heat fluxes, atmospheric state variables, soil variables, and meteorological observations.

In addition, the reference publication links provide detailed descriptions of individual sites, including ecosystem characteristics, vegetation information, terrain conditions, soil properties, instrumentation, observation protocols, and other site-specific information that are essential for data interpretation and scientific applications.

Rather than duplicating these heterogeneous datasets and metadata within the benchmark product itself, we chose to provide direct access to the original data sources and site publications. This approach ensures that users can obtain the most complete and up-to-date information available for each site, while maintaining full transparency and traceability of the underlying observations. It also appropriately acknowledges the contributions of the original site investigators and data providers whose long-term measurements form the foundation of this benchmark dataset.

To make this clearer, we have revised Section 5 and Appendix Table A2 to explicitly state that the provided links enable access to both site-level observational data and detailed site metadata, including ecosystem, terrain, and soil characteristics.

Detailed comments 63

Section 6. The conclusions section is a summary of results. The conclusions should be the "take home" message for the readers of the manuscript. The summary of results belongs in the abstract. Detailed results belong in the results section. I suggest that this section should be rewritten.

Author Respond: We thank the reviewer for this helpful comment and fully agree that the original conclusion section placed excessive emphasis on summarizing individual results and performance metrics that had already been presented in detail in the Results section.

Following the reviewer's suggestion, we completely rewrote Section 6. Rather than reiterating numerical results and model evaluation statistics, the revised conclusion now focuses on the main take-home messages of the study, namely: (1) the development of a continuous half-hourly latent heat flux benchmark dataset for China spanning 2000–2024, (2) the scientific value of overcoming data discontinuity and heterogeneous observation periods in flux tower records, (3) the broader applicability of the dataset for evapotranspiration product evaluation, land surface model benchmarking, and water–energy cycle research, and (4) the contribution of the proposed framework to enhancing the long-term usability of ChinaFlux observations.

The revised conclusion therefore emphasizes the significance, utility, and potential applications of the dataset rather than repeating detailed performance results. We believe this revision substantially improves the readability of the manuscript and aligns the conclusion section more closely with the purpose of an ESSD data paper.

The revised ***Conclusion***:

6 Conclusion

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

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

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

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