Response to reviewer #2 Comments - Round 2

Reviewer general comments: This manuscript has improved a great deal. Most of my concerns have been addressed. I have two remaining main comments and few detailed suggestions. I support publication of this important effort.

Author Respond: We sincerely appreciate your careful re-evaluation of our revised manuscript and your positive assessment of this dataset-oriented work, as well as your support for its publication. We are deeply grateful for your valuable and constructive suggestions throughout the review process. Your thoughtful comments have substantially enhanced the scientific soundness, logical clarity, readability, and overall rigour of this manuscript.

We have carefully examined each of your remaining major comments and detailed minor suggestions, and made every effort to implement corresponding revisions within the available revision time. All points raised have been addressed point-by-point in the following response-to-comments section, together with explicit indications of where modifications have been implemented within the revised manuscript. We hope that the revised version meets your expectations for publication in Earth System Science Data.

Major comments:

Detailed comments 1
1. I thank that authors for clarifying most of the uncertainties about gap-filled data, especially the 10 sites that are completely gap-filled with no indication of which data are original vs. gap-filled (I believe - I have one clarifying question about this in the detailed comments). I remain uncertain, however, about how data from the remaining 40 sites are handled. Do the remaining 40 sites include some investigator gap-filled data? If so, are these identified in some way? I could not find this information in the text. Ideally I would suggest that the authors eliminate all investigator gap-filled data and use their own gap-filling to make the data product as consistent as possible. I have questions related to this topic in the detailed comments. I strongly recommend that the authors clarify this point. My suggestion that they replace all investigator gap-filling with their own method is that - a suggestion. If the authors do retain some investigator gap-filled data I would strongly suggest that they label these points separately from the original observations, and separate from their own gap-filling.

Author Respond: Thank you for emphasizing this important point. We agree that the treatment of investigator gap-filled data must be clearly documented to ensure the internal consistency and transparency of the data product. We have revised the manuscript, Appendix Table A1, and the Readme.txt file to clarify this issue.

For the 40 sites with original flux tower observations, investigator gap-filled data were not retained. Before model development and dataset generation, we strictly removed all values that were identified as investigator gap-filled or otherwise not high-quality original observations according to the available QA/QC information. Only QA/QC-filtered original LE observations were retained as training and reference data. Therefore, for these 40 sites, the T flag in the released half-hourly dataset denotes original QA/QC-filtered tower observations, and missing values within the observation period were filled only by our AutoML framework and labeled as F. Temporally prolonged values beyond the observation period were generated by the same framework and labeled as P.

For the 10 marked sites, the situation is different. These sites were supplied as complete LE time series within their nominal observation periods, but the source records do not distinguish original observations from investigator gap-filled values. Therefore, these provider-supplied complete time series were retained as provided and labeled as T within the nominal observation period. We explicitly state that, for these 10 marked sites, T should not be interpreted as purely original observations, but rather as provider-supplied complete records with unknown original/gap-filled status. These sites are identified in Appendix Table A1 and in the Readme.txt file so that users can include or exclude them according to their research objectives.

Thus, no investigator gap-filled data are retained for the 40 original-observation sites. The only cases in which original and investigator gap-filled values cannot be separated are the 10 marked provider-supplied complete-series sites, and this limitation is explicitly labeled and documented in both the manuscript and the final data product.

The modified content in the ***Data availability*** section:

1) Half-hourly LE data (2000–2024). The half-hourly LE dataset represents the core product of this study and provides complete time series for all flux tower sites. **Each record is accompanied by a quality flag (QC_LE) indicating its data source: T denotes half-hourly LE derived from original flux tower observations; F represents gap-filled LE generated within the observation period using the AutoML framework; and P indicates LE values produced through temporal prolongation beyond the observation period using the same framework. For the 40 sites with original observations, T corresponds to original QA/QC-filtered tower observations. For the 10 marked sites, however, T represents**

provider-supplied complete LE time series within the nominal observation period and should not be interpreted as purely original observations. This distinction is explicitly stated in Appendix Table A1 and in the Readme.txt file. All half-hourly data follow a unified time format, with timestamps referenced to local site time to ensure temporal continuity and traceability. Timestamps are provided in the format YYYYMMDD–HHMM (e.g., 20120101–1530), representing local date and time at each site. Each timestamp denotes the beginning of the corresponding 30 min flux-averaging period. For example, the records for 1 January 2012 range from 20120101–0000 to 20120101–2330, representing 48 half-hourly time steps, and 20120101–1530 indicates the start of the 15:30–16:00 local-time interval.

Detailed comments 2

2. The description of methods has improved considerably. I have one remaining concern / suggestion. It can still be difficult at times to separate the workflow you used to test the gap filling and record prolongation algorithms from the workflow used for gap filling and record prolongation. These are mixed together in Figure 2. I believe it would help the clarity of the document and the long-term value of the manuscript if Figure 2 was used only to describe how the final data products are created. The authors could make a smaller "cut out" describing the work flow for algorithm testing. The results of the manuscript can then easily cite these two different figures and work flows and remove the remaining confusion.

Author Respond: We thank the reviewer for this helpful suggestion. We agree that the previous Figure 2 mixed the production workflow used to generate the final LE products with the evaluation workflows used to test the gap-filling and temporal prolongation algorithms, which could cause confusion.

To address this issue, we have revised Figure 2a so that it now only describes the workflow used to generate the final released data products. Specifically, the revised Figure 2a focuses on data acquisition, preprocessing, site-specific AutoML model training, gap-filling of actual missing values, temporal prolongation to 2000-2024, and generation of the half-hourly and aggregated LE products.

We have also added new figure to separately describe the algorithm evaluation workflows. Figure 2b shows the artificial-gap experiments used to evaluate the gap-filling algorithms, including the four gap-length scenarios and comparison with MDS, RF, and XGBoost. Figure 2c shows the forward/backward and training-length experiments used to evaluate the temporal prolongation algorithm.

We now explicitly state that the artificial-gap experiments and temporal-split experiments were used only for algorithm evaluation, whereas the final data products were generated by retraining the site-specific AutoML models using all available high-quality LE observations at each site.

The modified content in the ***Data and methodology*** section:

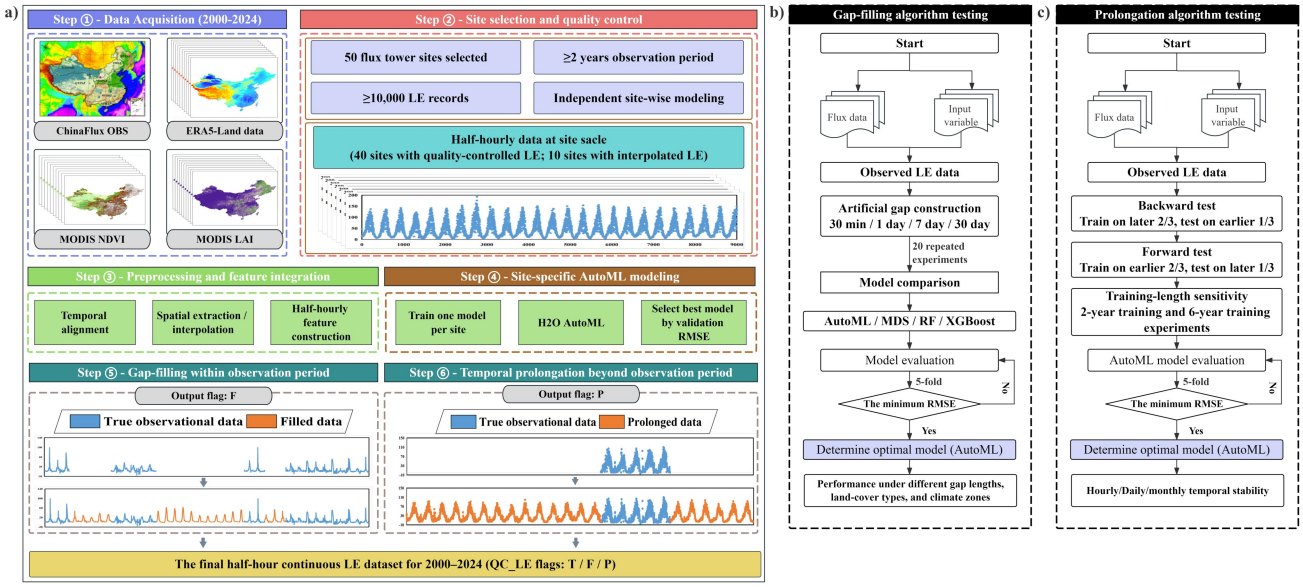


Figure 2: Workflow for generating the final half-hourly LE data products from 2000 to 2024 (a). Workflows used to test the gap-filling (b) and prolongation algorithms (c). The evaluation workflows were independent from the production workflow. Artificial gaps and forward/backward temporal splits were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

Detailed comments 3

My primary confusion remains in the AutoML algorithm vs. MDS, RF and XGBoost. The text says, if I understand it correctly, that AutoML selects from among gap-filling algorithms, suggesting that it uses one of these three methods for each site in its final gap filling. The text, however, compares AutoML to these algorithms. Please clarify this issue.

Author Respond: Thank you for pointing out this important source of confusion. We agree that the previous wording could be read as if AutoML were selecting one method from among MDS, RF, and XGBoost. That is not the case. We have revised the text to make the methodological structure explicit.

In this study, four methods are considered as separate and independent approaches: AutoML, RF, XGBoost, and MDS. AutoML is an automated framework that searches over its own internal candidate learners and hyperparameter configurations, and then selects the best-performing model based on validation RMSE. In contrast, RF, XGBoost, and MDS are standalone benchmark methods used only for comparison. They are trained and evaluated independently and do not form part of the AutoML search space.

We also clarified that the benchmark methods were used only within the artificial-gap evaluation framework, whereas the final released gap-filled LE dataset was generated using the site-specific AutoML model. This revision should remove the ambiguity between the AutoML framework and the benchmark methods and make the comparison structure internally consistent.

The modified content in the ***Data and methodology*** section:

To facilitate comparative evaluation of different gap-filling approaches, **four independent methods were considered in this study: AutoML, random forest (RF), extreme gradient boosting (XGBoost), and marginal distribution sampling (MDS). AutoML was treated as an automated modeling framework that searches over multiple internal candidate algorithms and hyperparameter configurations and then selects the best-performing leader model based on validation RMSE; it does not select from among the benchmark methods used for comparison in this study.** The traditional marginal

distribution sampling (MDS) method (Foltýnová et al., 2020) and two commonly used machine learning algorithms, RF (Khan et al., 2021) and XGBoost (Liu et al., 2025), were implemented as benchmark methods. **These three benchmark methods were trained and evaluated independently of AutoML, and therefore do not belong to the AutoML framework.** These benchmark methods were also applied only within the artificial-gap evaluation framework and were not used to generate the final released gap-filled LE dataset. Model performance was assessed using five-fold cross-validation during the training stage to optimize model configuration and reduce the risk of overfitting. After model selection, the optimal model for each site was retrained using all available high-quality LE observations and then applied to fill the missing LE values. **For the final released dataset, the site-specific AutoML model was used for production, rather than MDS, RF, or XGBoost.** Consequently, a site-specific gap-filling model was established for each site, providing a reliable basis for subsequent temporal prolongation of LE time series.

Specific comments:

Detailed comments 1
line 19-20. **"an automated machine learning approach (AutoML) of H2O framework"** **The grammar needs work here. I don't think the addition "of H2O framework" helps the text.**

Author Respond: Thank you for this helpful comment. We agree that the phrase "of H2O framework" was awkward and unnecessary. We have revised the sentence to improve grammatical correctness and readability by removing this part and simplifying the expression to "an automated machine learning approach (AutoML)." The revised sentence now reads more smoothly and clearly conveys the methodological framework used in this study.

The modified content in the ***Abstract*** section:

The framework is built upon an automated machine learning approach (AutoML) and integrates ERA5-Land reanalysis data with MODIS vegetation indices, enabling accurate gap-filling within observation periods and reliable prolongation beyond observation intervals.

Detailed comments 2
Line 65 begins an exceedingly long paragraph. Please break this up into paragraphs according to the flow of ideas within this text. Perhaps this is already broken into smaller paragraphs but I can't tell due to the typesetting.

Author Respond: Thank you for this helpful suggestion. We agree that the paragraph appeared excessively long in the typeset version and could reduce readability. In fact, the original manuscript already contained a paragraph break before "In China," but this separation was not sufficiently clear in the formatted layout. To address this issue, we have inserted blank lines between paragraphs throughout the Introduction to make the paragraph boundaries explicit and improve the visual clarity and readability of the text. No substantive changes were made to the content.

The modified content in the ***Introduction*** section:

Paragraph 1: Within the existing ground observation systems, the eddy covariance (EC) technique has been widely regarded as the most important data source for constraining and evaluating ET models, as it directly measures turbulent exchanges between the land surface and the atmosphere. EC observations provide half-hourly measurements of latent heat flux (LE), while the associated meteorological or environmental variables are measured concurrently at the flux tower sites, thereby serving as an essential ground-based benchmark for evapotranspiration research. However, EC flux time series are commonly affected by substantial data gaps and long periods of missing observations due to instrument maintenance, complex meteorological conditions, and data transmission interruptions, with large variations in effective observation lengths among sites. In addition, EC measurements are subject to systematic uncertainties associated with energy balance non-closure, which may affect the accuracy of flux estimates even when observations are available. Taking the FLUXNET2015 dataset as an example, the average missing rate of hourly LE data is approximately 40 %, exceeding 70 % at some sites, and long gaps lasting more than 30 d account for a considerable fraction of the missing records (Foltýnová et al., 2020; Zhu et al., 2022). Although gap-filling is necessary to generate continuous LE time series, short gaps are often handled using various statistical or empirical approaches. Nevertheless, accurately reconstructing long and continuous LE time series remains challenging because evapotranspiration is governed by complex interactions among atmospheric conditions, vegetation dynamics, and water availability. Moreover, the

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

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

Detailed comments 3

Line 195.

"for each flux towers"

each flux tower.

Author Respond: Thank you for pointing out this grammatical error. We have corrected "for each flux towers" to "for each flux tower" to ensure proper singular–plural agreement. The revised sentence is now grammatically correct and clearer.

The modified content in the ***Data and methodology*** section:

Detailed information on site locations, underlying surface types, climate zone classifications, and observation periods is provided in Table A1, and data source information **for each flux tower** can be seen in Table A2.

Detailed comments 4

Thank you for pointing out the citations for the data sets in Table A2. This is traceable and acceptable.

Author Respond: Thank you for your kind comment. We are pleased that the citations for the datasets in Table A2 are considered traceable and acceptable.

Detailed comments 5

Line 201-202.

"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."

Thank you for addressing this important issue. I have two remaining questions that I suggest that you should clarify in the text. First, for the 10 sites that provide continuous time series, is it true that those sites do not distinguish the original observations from the gap-filled data points? If so, please make this clear. If not, you could do an independent gap filling. I believe this would improve the consistency of your data set.

Second, are there any investigator gap-filled data in the records from the other 40 sites? If so, do you retain those? Are they identified in the data sets? I would suggest that you replace all investigator gap-filled data with your gap-filling - if this is possible. I ask this because I note, for example, that sites 12-19 have less than 10% missing data. This is remarkably complete data if there was no gap filling.

I wonder if a flow chart would help to explain these issues of gap filled (investigator vs. yours) vs. original data, and data labelling in your data base?

Author Respond: Thank you for raising these important questions. We have revised the text to clarify the data provenance and labeling strategy. First, for the 10 sites that provide continuous LE time series, the source records do not distinguish original observations from investigator gap-filled values. Therefore, these records were retained exactly as provided by the data contributors, and we clearly state this limitation in the revised manuscript. Second, for the other 40 sites, no investigator gap-filled data are retained in our dataset. We had already strictly removed all such values before our processing, so the records used in this study consist only of original observations, with missing values filled subsequently by our own method.

We also agree that clear data labeling is essential for transparency and reproducibility. Accordingly, we further clarified in the manuscript that all sites are processed independently, that the 10 continuous-series sites are explicitly identified in Table A1 and the Readme file, and that the final dataset uses separate files and flags to distinguish original, filled, and prolonged data. We did not add a flow chart, as we believe the revised text and dataset documentation already provide a sufficiently clear explanation of the distinction between investigator-provided continuous series and our own gap-filling outputs.

The modified content in the ***Data and methodology*** section:

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

Readme.txt file, allowing users to include or exclude them according to their specific research needs. Given the overall scarcity of ChinaFlux data, retaining these sites provides users with greater flexibility in selecting data for their specific research objectives.

Detailed comments 6

Figure 1a. This background is helpful and I can readily identify the flux site vegetation types. I have two remaining suggestions. First, please define the vegetation cover categories and the source of that land cover data. Second, please delete the elevation color bar that remains in the figure.

Author Respond: Thank you for this helpful suggestion. We have revised Fig. 1a and the corresponding text accordingly. First, we added a clear description of the vegetation cover categories and specified that they were classified based on the IGBP land cover scheme. Second, we identified the land cover background as the MODIS land cover product (MCD12Q1) and cited Friedl et al. (2010) as the source reference. In addition, we cross-validated the land cover information with the site-provided metadata to ensure consistency with the actual vegetation types at each flux tower. Finally, we removed the elevation color bar from Fig. 1a as suggested.

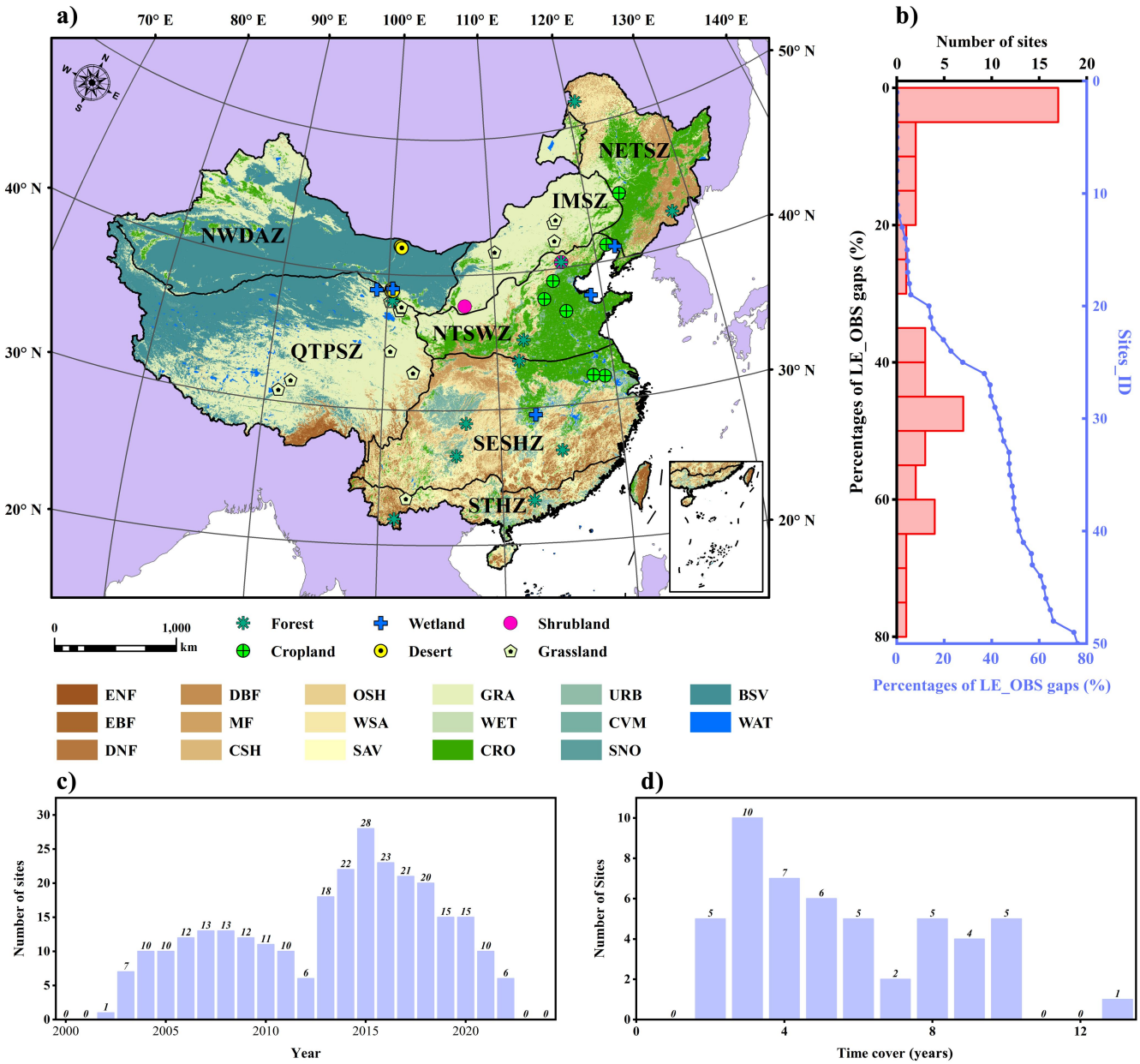
The modified content in the ***Data and methodology*** section:

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

The added references are as follows:

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

The modified ***Figure1***:



Detailed comments 7

Lines 275-277.

"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."

Please delete the second sentence. It is redundant.

More importantly, this again raises my earlier question about investigator gap-filled data and the "Number of LE Data" and "Missing Ratio" columns of Table A1. It would make the most sense to me if these columns reflected the number of "reliable records" used in this data product. Is that what is reported in Table A1? If not, I strongly suggest adding these additional columns somewhere in this document. I also strongly suggest clarifying, in your final data product, the origins of each data point - original observation, investigator gap filled, unknown (if you can't distinguish original observations from investigator gap filling) and gap filled using your methods. If possible, I also suggest removing all

investigator gap-filled data to provide the most internally consistent data product possible. If you choose to retain investigator gap filling, please make this clear in the text and in the final data product.

Author Respond: Thank you for this important comment. We have deleted the redundant second sentence as suggested. We also clarified in the revised text that the "Number of LE Data" and "Missing Ratio" columns in Table A1 are based on the quality-screened reliable records retained for analysis, and therefore reflect the records used in model development. In addition, the final data product already provides explicit data-source labels for each half-hourly record through the QC_LE flag system (T/F/P), and the distinction between the 40 sites with QA/QC-filtered original observations and the 10 sites with provider-supplied continuous time series is clearly documented in Table A1 and the Readme.txt file. For the 40 original-observation sites, no investigator gap-filled values were retained prior to our processing. For the 10 provider-supplied sites, the source records do not allow separation of original and investigator gap-filled values, and this limitation is explicitly stated in the manuscript and dataset documentation.

Detailed comments 8

Figure 2 is a hybrid. It shows both the research the investigators used to identify (actually justify, since it appears that AutoML was selected at the onset of the study) their favored gap-filling method and the methods used to fill the final data set. I suggest that this should be broken into two diagrams. One can outline the testing of methods and a second could outline the final data gap filling only with the selected algorithms. I believe this would make the data processing much clearer to most readers.

Author Respond: We thank the reviewer for this helpful suggestion. We agree that the previous Figure 2 mixed the production workflow used to generate the final LE products with the evaluation workflows used to test the gap-filling and temporal prolongation algorithms, which could cause confusion.

To address this issue, we have revised Figure 2a so that it now only describes the workflow used to generate the final released data products. Specifically, the revised Figure 2a focuses on data acquisition, preprocessing, site-specific AutoML model training, gap-filling of actual missing values, temporal prolongation to 2000-2024, and generation of the half-hourly and aggregated LE products.

We have also added new figure to separately describe the algorithm evaluation workflows. Figure 2b shows the artificial-gap experiments used to evaluate the gap-filling algorithms, including the four gap-length scenarios and comparison with MDS, RF, and XGBoost. Figure 2c shows the forward/backward and training-length experiments used to evaluate the temporal prolongation algorithm.

We now explicitly state that the artificial-gap experiments and temporal-split experiments were used only for algorithm evaluation, whereas the final data products were generated by retraining the site-specific AutoML models using all available high-quality LE observations at each site.

The modified content in the ***Data and methodology*** section:

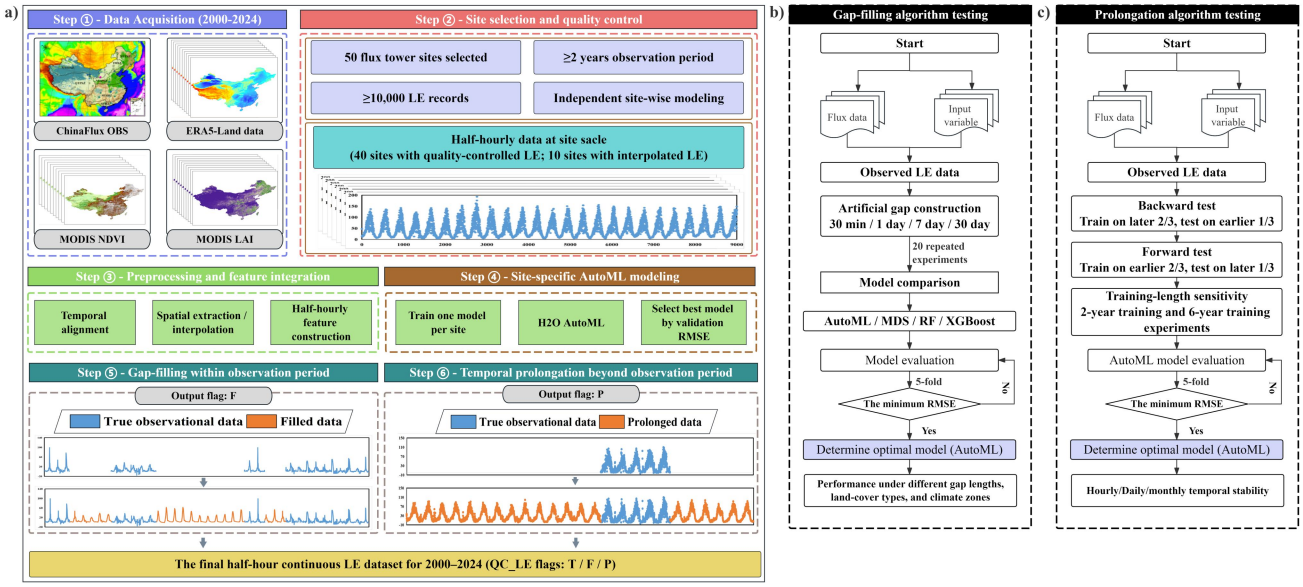


Figure 2: Workflow for generating the final half-hourly LE data products from 2000 to 2024 (a). Workflows used to test the gap-filling (b) and prolongation algorithms (c). The evaluation workflows were independent from the production workflow. Artificial gaps and forward/backward temporal splits were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

Detailed comments 9

Lines 287-289.

"AutoML of H2O framework (LeDell and Poirier, 2020; H2O.ai, 2023) was adopted as the core modeling tool for LE gapfilling 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)."

The phrase "which has been widely used" is a dangling modifier. What has been widely used? I cannot tell.

Author Respond: Thank you for pointing out this grammatical issue. We agree that the phrase "which has been widely used" created an unclear reference in the original sentence. To avoid the dangling modifier and improve readability, we revised the sentence by explicitly identifying the subject. Specifically, we changed "AutoML of H2O framework" to "The H2O AutoML framework" and split the sentence into two sentences, with "H2O AutoML" clearly serving as the subject of the statement about its use in environmental and geoscientific applications. The revised text is now grammatically clearer and more precise.

The modified content in the ***Data and methodology*** section:

2.4.1 AutoML

AutoML of H2O framework (LeDell and Poirier, 2020; H2O.ai, 2023) was adopted as the core modeling tool for LE gap-filling in this study. **H2O AutoML has been widely used** for regression and prediction tasks in environmental and geoscientific applications (Guo et al., 2024; Li et al., 2025b; Zhao et al., 2026). Unlike conventional machine learning approaches that require manual specification of model types and hyperparameters—often leading to high tuning costs and reduced transferability across sites—AutoML automatically performs model selection, hyperparameter optimization, and model ranking within a predefined search space (LeDell and Poirier, 2020).

Detailed comments 10

Lines 300-311. Is one algorithm chosen for each site? Or can AutoML change algorithms (over time?) within the record of a single flux tower? This text suggests that only one algorithm is used for each tower - if I understand the text correctly.

Author Respond: Thank you for this helpful question. We have revised the text to clarify how AutoML was used for each flux tower site. For each site, two site-specific final models were selected from the H2O AutoML leaderboard: one model was used for gap-filling within the observation period, and the other was used for temporal prolongation beyond the observation period. Each model was selected separately for its corresponding task based on the lowest validation RMSE among the evaluated algorithms and hyperparameter configurations. After selection, the models were fixed for the given site and task, and the selected algorithms did not change over time within the record of a given flux tower. We have added this clarification to avoid the possible misunderstanding that AutoML dynamically changes algorithms within a single time series.

The modified content in the ***Data and methodology*** section:

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. **For each flux tower site, two site-specific final models were selected from the H2O AutoML leaderboard: one model for gap-filling within the observation period and another model for temporal prolongation beyond the observation period. Each final model corresponded to the candidate model with the lowest validation RMSE among all evaluated algorithms and hyperparameter configurations for its respective task. After selection, these two models were fixed for the corresponding site and task, and the selected algorithms did not change over time within the record of a given flux tower.** In this study, no strict limit was imposed on the number of candidate models (i.e., the number of models is adaptively determined by the AutoML process), allowing the framework to fully explore the model space and select the optimal model for each flux tower site.

Detailed comments 11

Lines 300-311 seem to conflict with Step 2 of Figure 2 which suggests that AutoML is one of the possible models to be used. Instead, this text suggests that AutoML selects from among a set of algorithms. This is confusing. Please clarify and make Figure 2 and this text internally consistent.

Author Respond: Thank you for pointing out this inconsistency between the text and Fig. 2. We agree that the previous presentation could give the impression that AutoML was treated as one model among several candidate algorithms, whereas the text describes AutoML as a framework that searches across multiple algorithms and hyperparameter configurations. We have revised the manuscript to clarify this workflow.

In the revised text, we now explicitly state that AutoML was not treated as a single predefined algorithm, but as an automated modeling framework. The benchmark methods, including RF and XGBoost, were also optimized independently by testing different parameter combinations and selecting the parameter setting with the lowest validation RMSE. After the optimized AutoML and benchmark models were obtained, their performances were compared, and the best-performing approach was selected for generating the final data product. The results showed that AutoML performed best or comparably best in almost all

evaluation scenarios; therefore, AutoML was ultimately adopted for producing the gap-filled and temporally prolonged LE time series.

We also revised Fig. 2 to make it internally consistent with the text. Specifically, the figure now distinguishes the candidate modeling approaches, validation-based model optimization, performance comparison, and final selection for data production. This revision avoids implying that AutoML is a single algorithm equivalent to RF or XGBoost and clarifies that AutoML is an algorithm-selection and hyperparameter-optimization framework.

The modified content in the ***Data and methodology*** section:

During the AutoML search process, multiple commonly used regression algorithms, including tree-based models and their ensemble variants, were evaluated, and the optimal model was automatically selected based on validation performance. **Here, AutoML was not treated as a single predefined algorithm, but as an automated modeling framework that searches across multiple candidate algorithms and hyperparameter configurations. In parallel, the benchmark machine learning models were also tuned using the same validation-based principle: different parameter combinations were iteratively tested, and the parameter setting with the lowest validation root mean square error (RMSE) was retained as the optimized model for that method. After the optimized AutoML and benchmark models were determined, their performance was compared, and the best-performing modeling approach was selected for generating the final data product. Because AutoML consistently showed superior or comparable performance in almost all evaluation scenarios, it was ultimately adopted for producing the gap-filled and prolonged LE time series.**

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

Detailed comments 12

Lines 316-317.

"…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."

As with the algorithms, this causes confusion. I suggest that Figure 2 be split up. Please illustrate the testing of algorithms from the final data processing used to create your data set. I think that the "testing" diagram could be a simpler subset of the full path shown in Figure 2. Figure 2 could then be simplified by removing the algorithm testing elements (gap lengths?) that are not part of the final data processing.

Author Respond: We appreciate this constructive comment. We agree that mixing algorithm-testing components and final dataset-production workflows within one single figure leads to potential misinterpretation, as the artificial-gap experimental setups are only used for algorithm benchmarking and are not involved in generating the released dataset.

Following your suggestion, we have redesigned and split original Figure 2 into three sub-panels to clearly separate the final data-generation workflow from algorithm evaluation workflows:

Figure 2a: presents the complete workflow for producing the final released half-hourly LE dataset (production pipeline only, all artificial-gap testing elements are removed from this panel).

Figure 2b: illustrates the testing workflow for the in-period gap-filling algorithm, including the artificial gap scenarios (30 min, 1 d, 7 d, 30 d), multi-method inter-comparison and performance assessment. This represents the algorithm-validation subset.

Figure 2c: shows the testing workflow for the temporal prolongation algorithm, containing forward-/backward extrapolation validation and temporal-stability evaluation.

Accordingly, we have revised the corresponding caption and main-text descriptions to explicitly distinguish production procedures from experimental evaluation procedures. All related text passages have been updated to match the revised figure structure in the revised manuscript.

Detailed comments 13

Line 361.

"And RMSE was used…"

Please delete "And".

Author Respond: Thank you for pointing this out. We have deleted "And" at the beginning of the sentence as suggested. The revised sentence now reads more smoothly and formally.

Detailed comments 14

Line 362-363.

"Except to these quantitative metrics, a qualitative assessment of diel-cycle reconstruction was conducted for representative sites."

I'm sorry, but I don't understand this sentence.

Author Respond: Thank you for pointing out this unclear sentence. We agree that the original expression "Except to these quantitative metrics" was grammatically incorrect and did not clearly convey the purpose of the assessment. We have revised the sentence to clarify that, in addition to the quantitative metrics, we also evaluated the temporal behavior of the reconstructed LE time series at representative sites, as illustrated in Figs. 5, 6, and 9. This comparison was used to assess whether the reconstructed time series preserved the observed diurnal patterns, amplitude variations, and temporal continuity during extended missing periods.

The modified content in the ***Data and methodology*** section:

RMSE was used as the model selection criterion, whereas CC and PBias were reported for supplementary performance evaluation. **In addition to these quantitative metrics, we further evaluated the temporal behavior of the reconstructed LE**

time series at representative sites. This assessment examined whether the reconstructed LE time series preserved the observed diurnal patterns, amplitude variations, and temporal continuity during extended missing periods.

Detailed comments 15

Line 405-406.

"The flag T denotes half-hourly LE values directly derived from original flux tower observations."

Are these all observations for the 40 sites that were not gap-filled without labels for the gap-filled times? Or are investigator gap-filled data points included under the "T" designation? See my earlier suggestions.

Author Respond: Thank you for raising this important clarification. The meaning of the T flag depends on the type of site, and we have clarified this distinction in Section 5 ("Data availability"), Appendix Table A1, and the Readme.txt file.

For the 40 sites with original flux tower observations, T denotes half-hourly LE values derived from original QA/QC-filtered tower observations. Investigator gap-filled values were not retained for these 40 sites; missing values within the observation period were filled only using our AutoML framework and are labeled as F.

For the 10 marked sites, however, the source records provide only complete LE time series within the nominal observation period, and the original observations cannot be distinguished from investigator gap-filled values. Therefore, for these 10 sites, T represents the provider-supplied complete LE time series and should not be interpreted as purely original observations. This limitation and distinction are explicitly documented in Appendix Table A1 and in the Readme.txt file, allowing users to identify these sites and include or exclude them according to their research needs.

Thus, investigator gap-filled data are not included under the T designation for the 40 original-observation sites, but the T flag for the 10 marked sites represents provider-supplied complete records with mixed or unknown original/gap-filled status.

Detailed comments 16

Section 3.1.1. My understanding of AutoML, based on this manuscript, is that it selects from among the algorithms that it is being compared to here - MDS, RF and XGBoost. How can it be that it performs better than all of these, if in fact AutoML selects one of these for each site? I am afraid that I do not understand the text describing how AutoML works. Can you please clarify this text?

Author Respond: Thank you for raising this important point. We agree that the previous description could lead to confusion about the relationship between H2O AutoML and the benchmark methods. We have revised Section 3.1.1 to clarify this issue.

H2O AutoML was not used to select from the benchmark methods compared in this study. In particular, MDS is not part of the H2O AutoML candidate algorithm set; it is an independent conventional gap-filling method used as a benchmark. The standalone RF and XGBoost models used for comparison were also implemented and tuned independently outside the AutoML framework. By contrast, H2O AutoML searches across a broader model space, including GBM, DRF, XGBoost, GLM, DNN, and stacked ensemble models, together with multiple hyperparameter configurations. Therefore, the final AutoML leader model may be a stacked ensemble or another optimized learner rather than being identical to the standalone RF or XGBoost benchmark model.

This distinction explains why AutoML can outperform the benchmark methods. AutoML benefits from a broader algorithmic search space, automated hyperparameter optimization, model ranking based on validation performance, and the possible use of stacked ensembles. After all optimized methods were evaluated, AutoML showed superior or comparable performance in

almost all evaluation scenarios and was therefore selected for producing the final gap-filled and temporally prolonged LE data product. We have revised the text to make this workflow explicit and to avoid the misunderstanding that AutoML simply selects among MDS, RF, and XGBoost.

Detailed comments 17

Figure 3. The caption is improved. But I'm sorry, I still don't know what is being reported in the figure. The caption says, 'for each flux tower'. But there is only one figure which I believe(?) describes all of the flux towers. There is also the mention of 20 repeated experiments - which I don't understand. Please explain, in the caption, in good detail, the populations that are used to create the figure. Hopefully this is mirrored in the methods. After this you can say "as in Figure 3" for any additional descriptions of the same populations. My apologies if I am slow to follow your work, but this needs to be clear to typical readers.

Author Respond: Thank you for this helpful comment. We agree that the previous caption did not sufficiently explain the population represented by each boxplot in Fig. 3, especially the relationship among the 50 flux tower sites, the 20 repeated experiments, and the displayed distributions.

We have revised the caption to clarify that each boxplot represents the distribution of site-level performance across the 50 flux tower sites for a given gap-filling method and artificial gap scenario. For each site, artificial gaps were generated and filled in 20 repeated experiments. The performance metrics were first calculated for each repetition and then averaged across the 20 repetitions, so that each site contributes one averaged value for each metric. Therefore, each boxplot in Fig. 3 is constructed from 50 site-level averaged values. We also added explanations of the boxplot elements, including the interquartile range, median, whiskers, mean point, and the numerical mean values shown in the figure.

To make this definition clear in the methods as well, we added corresponding text in Section 3.1.1 after the description of the artificial gap experiments. This added text explains how the 20 repeated experiments were summarized at the site level and how the 50 site-level values were used to construct the boxplots in Fig. 3. This revision should make the figure easier to interpret and ensure consistency between the methods and figure caption.

The modified content in the ***Data and methodology*** section:

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. **For each site, artificial gap scenario, and gap-filling method, the performance metrics were first calculated separately for each of the 20 repeated experiments and then averaged across the 20 repetitions to obtain one site-level value for each metric. Therefore, in the cross-site comparison shown in Fig. 3, each boxplot is constructed from 50 site-level averaged values, with each flux tower contributing one value.**

The modified content in the ***Figure3***:

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

Detailed comments 18

I am confused by the caption of Figure 4. Figure 4 is a comparison of different gap length scenarios and gap filling methods. The caption, however, mentions "the final released dataset." This figure does not deal in any way with the "final released dataset", does it? I know that you are trying to state that this is methods evaluation and not a presentation of the final data set. I believe you have made this clear in the text already. I think this last sentence in the figure caption is confusing.

Author Respond: Thank you for pointing this out. We agree that the last sentence of the previous caption could be confusing because Fig. 4 presents only the method-evaluation results under different artificial gap-length scenarios and does not directly describe the final released dataset. We have therefore deleted the sentence referring to the final released dataset from the caption. The revised caption now focuses only on the purpose of Fig. 4, namely the comparison of gap-filling performance across methods, underlying surface types, climate zones, and artificial gap scenarios. This revision avoids implying that Fig. 4 represents or summarizes the final data product.

The modified content in the ***Figure4***:

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.

Detailed comments 19

Line 719-720.

"Timestamps are provided in the format YYYYMMDD–HHMM (e.g., 20120101–1530), representing local date and time at each site."

Does 1530 mark the beginning, middle or end of the 30 minute period of flux data? Please make sure this is clear in the metadata for the data product.

Author Respond: Thank you for this helpful comment. We have revised the manuscript and metadata description to clarify the meaning of the half-hourly timestamps. Each timestamp denotes the beginning of the corresponding 30 min flux-averaging period. For example, records for 1 January 2012 range from 20120101–0000 to 20120101–2330, corresponding to 48 half-hourly time steps, and 20120101–1530 represents the start of the 15:30–16:00 local-time interval. This clarification has also been added to the metadata of the data product to ensure that users can correctly interpret the temporal reference of each record.

The modified content in the ***Data availability*** section:

All half-hourly data follow a unified time format, with timestamps referenced to local site time to ensure temporal continuity and traceability. Timestamps are provided in the format YYYYMMDD–HHMM (e.g., 20120101–1530), representing local date and time at each site. Each timestamp denotes the beginning of the corresponding 30 min flux-averaging period. For example, the records for 1 January 2012 range from 20120101–0000 to 20120101–2330, representing 48 half-hourly time steps, and 20120101–1530 indicates the start of the 15:30–16:00 local-time interval.