Referee 1:

General comment:

The research article discusses the development of a benchmark dataset for global evapotranspiration (ET) estimation, addressing limitations in existing latent heat flux (LE) data from the FLUXNET2015 dataset. Current datasets suffer from short observation periods and significant data gaps, hindering climate change analysis and model validation. To overcome these challenges, the authors created a gap-filling and prolongation framework that generates seamless half-hourly and daily LE data from 2000 to 2022 across 64 sites. They employed a novel bias-corrected random forest algorithm for improved data accuracy, achieving a median RMSE of 32.84 W/m² for hourly and 16.58 W/m² for daily data. The resulting dataset enhances ET modeling, water-carbon cycle monitoring, and climate change research.

The study is one of the pioneering efforts to utilize a bias-corrected random forest approach to enhance data gap-filling performance. I suggest minor revisions to address some specific questions before proceeding with publication.

Reply: Thank you for taking your precious time to review our article and also giving us such a positive comment. We have made a detailed reply and they are as follows. We have also carefully examined the entire manuscript and corrected some ambiguous or incorrect expressions therein. All the changes are marked in red.

Detailed comments:

1: Figure 3 - From the diagram there are two RF models being trained and evaluated. Please indicate that LE and Bias without single quote serve as observational ground-truth labels in Model training box.

Reply: Thank you for your suggestion. We have added the explanation of LE and Bias without single quote in the caption of Figure 3.

> ***In revised paper Line 194-195:***
> *"LE and Bias with single quotes indicate predicted values, whereas those without single quotes indicate ground observations."*

2: In Model validation box, there is only predicted values instead of true values being indicated. Please add that true LE and Bias are used to evaluate the performance of RF1 and RF2 and indicate performance metrics used for each model validation.

Reply: Thank you for your suggestion. According to your opinion, we have made corresponding modifications in Figure 3.

Firstly, we have changed the "Model validation" box to the "Model test" box to avoid misunderstanding.

Secondly, for RF1, we added true LE (LE_{test}) and used RMSE, CC, Bias as the performance metrics to evaluate the performance of RF1 (the original RF); for RF2, our aim is to accurately calculate and evaluate the predicted LE rather than Bias, so we directly evaluated the performance of RF1+RF2 ($LE'_{Bias\_corrected\ test}$ versus LE_{test}) and also used RMSE, CC, Bias as the performance metrics. We have added all the information into the "Model test box".

In revised paper Figure 3:

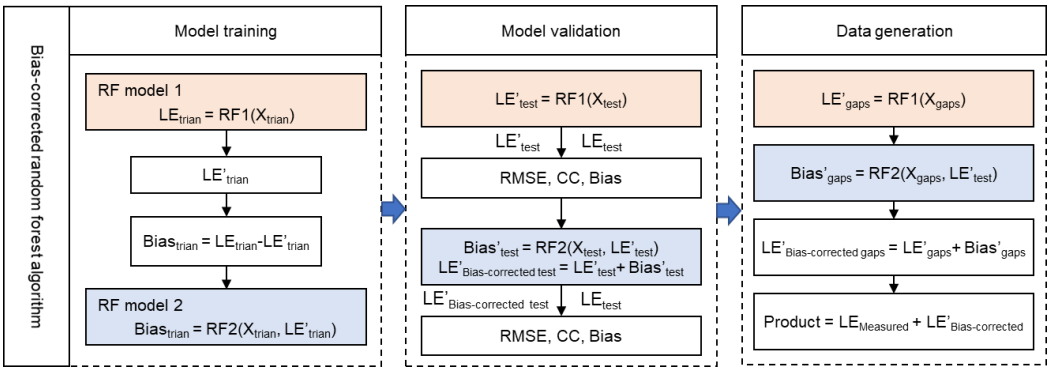

Figure 3 Schematic diagram of the bias-corrected RF algorithm. Train in the subscript indicates the training data. Test in the subscript indicates the test data. Gaps in the subscript indicates the data gap to be filled. LE and Bias with single quotes indicate predicted values, whereas those without single quotes indicate ground observations. X indicates the reference variables, including TA, WS, RH, PA, SW_IN, LW_IN, and NDVI. Prolonging daily data also has the same processing steps.

3: Figure 4 – It is hard to conclude Bias-corrected RF has better performance than the other two approaches as the mean values of RMSE of those three are tightly close to each other shown in the figure. Consider adding data labels to the mean RMSE values in the figure to highlight the findings. Same for Figure 5.

Reply: Thank you for your suggestion. Following your suggestions, we have added data labels of mean RMSE, mean CC, and mean Bias for "All sites" in the first line of Figure 4 and Figure 5. We calculated the average of the four scenarios as the result for the "ALL" scenario and use mean values of RMSE, CC, and Bias to replace the median values. We have updated the figure and also revised the corresponding value in the manuscript.

Overall, the bias-corrected RF can handle both short and long gaps very well. Specifically, it demonstrates significantly superior performance compared to the MDS method when handling very-long gaps. Additionally, it allows for temporal prolongation, which the MDS method cannot achieve. When compared with original RF, our method can improve the poor performance of RF under very-short and short gap scenario.

In revised paper Figure 4, Figure 5:

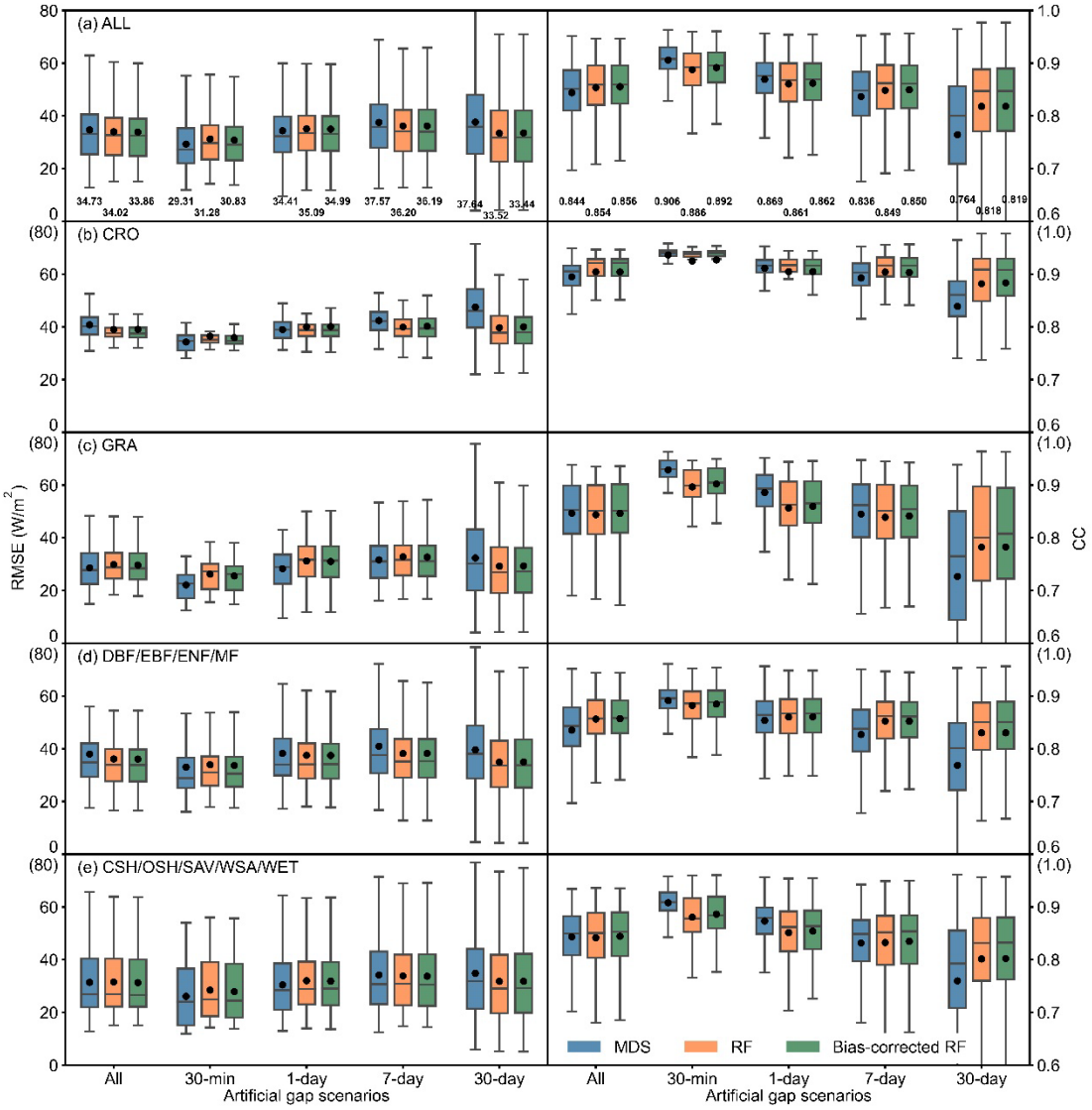

Figure 4 *The gap-filling performance of three algorithms under different gap-length scenarios. The left panels show the results of the root mean square error (RMSE, W/m^2) and the right panels show results of correlation coefficient (CC) between gap-filled values and observations. Different rows of this figure indicate different land cover types. The three horizontal lines of the boxes indicate the first quartile, median, and third quartile, respectively, and the black dots indicate the means. Data labels in this figure are the mean value of RMSE and CC. MDS: marginal distribution sampling. RF: random forest.*

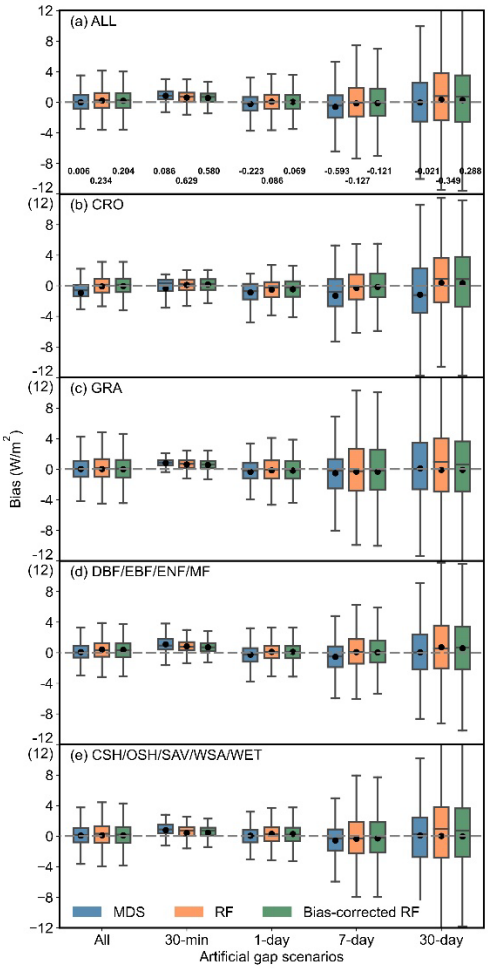

Figure 5 *The bias between gap-filled values and observations of three methods under different gap-length scenarios. Different rows of this figure indicate different land cover types. The three horizontal lines of the boxes indicate the first quartile, median, and third quartile, respectively, and the black dots indicate the means. Data labels in this figure are the mean value of Bias. MDS: marginal distribution sampling. RF: random forest.*

4: Line 185 – Please elaborate more on how you choose the best hyperparameters from 64 models. 64 models with 64 sets of parameters are obtained. For the sites with similar land type, are those models combined into one unified model by taking averages of parameters or still using different sets of parameters? Please explain it in more details.

Reply: Thank you for your suggestion.

For each site, the training and test dataset were generated 20 times, so we did the 10-fold cross validation for 20 times and gained 20 hyperparameter combinations. We found that for each site, the 20 hyperparameter combinations are almost the same. Therefore, we choose the hyperparameter combination based on two criteria: (1) achieving optimal model performance, and (2) exhibiting the highest frequency of occurrence across 20 experimental trials. **Consequently, each site has a site-specific and unique hyperparameter combination.**

In general, we trained model and get the best hyperparameter combination **site by site**. We did not unify the models from sites with the same land cover type. We have modified our expression in our revised paper to make it clearer.

In revised paper Line 185-191:

"For each site, the training and test dataset were generated 20 times, so we did the 10-fold cross validation for 20 times and gained 20 hyperparameter combinations. We found that for each site, the 20 hyperparameter combinations are almost the same. Therefore, we choose the hyperparameter combination based on two criteria: (1) achieving optimal model performance, and (2) exhibiting the highest frequency of occurrence across 20 experimental trials. Consequently, each site has a site-specific and unique hyperparameter combination."

5: In discussion section, please add potential limitations from this study in terms of variable importance, sensitivity and stability.

Reply: Thank you for your suggestion. We have added a "Advantages and disadvantages" part in discussion section, including the potential limitations from this study in terms of variable importance, sensitivity and stability.

In revised paper Line 416-427:

*"**5.3 Advantages and disadvantages***

Our study presents several notable advantages: 1) The bias-corrected RF shows better performance than the official MDS approach, especially for filling very long gaps (up to 30 days). Additionally, it allows for temporal prolongation, which the MDS method cannot achieve. Furthermore, our method enables the incorporation of a broader range of reference variables to establish a more robust non-linear relationship between LE and its drivers; 2) Compared to the FLUXNET2015 dataset, our hourly gap-filled data show improved quality and simpler implementation. The daily prolonged data provide extended temporal coverage (2000-2022) that is particularly valuable for evapotranspiration (ET) modeling and global-scale studies. However, some limitations in terms of variable importance, sensitivity and stability merit further discussion. The variable importance analysis (Section 5.2) indicated our method exhibits strong sensitivity to SW_IN data for gap-filling and to NDVI for prolongation. While we implemented bias correction between ground observations and ERA5-Land data, potential quality issues in SW_IN and NDVI inputs may still affect final results. Future improvements could incorporate higher-quality input data with more stable biases to enhance result reliability."