Response to referee comments: Anonymous Referee #2

We greatly appreciate the reviewers for dedicating their time and offering valuable suggestions to enhance the manuscript. We have carefully revised the manuscript following the reviewers' comments. For clarity, our point-by-point responses are listed below in blue text.

This study uses a Classifier-Guided Ensemble model to integrate process-based algorithm, ML-based model, and Hybrid model to identify the best model for a given case and subsequently estimate evapotranspiration (ET). The results demonstrate that the selected dominant model outperforms the individual constituent models. And the contribution of input parameters is also analyzed, which identifies the key drivers across all 3 model types. Overall, this study offers a tool to select the better model for a specific location among different models. Here, I have some comments and suggestions for the authors to consider in improving the manuscript, as well as some questions:

Response: Thank you for your constructive comments and suggestions. We have carefully revised the manuscript to address these points and improve the overall quality of the paper.

1. Line 142: The manuscript mentioned "ET as the sum of daytime and nighttime components", but the calculation methods for these two components are not clearly described. Please provide more details on how daytime and nighttime ET are estimated and combined.

Response: Thank you for your suggestion. We have revised Section 2.3 to explicitly describe the methodology for distinguishing and calculating daytime and nighttime ET components, following the methodology proposed by Mu et al. (2007, 2011). The updated text in Line 141 paragraph is as follows:

> Additionally, they improved the method to estimate vegetation cover fraction, soil heat flux, and parameters such as r_s, r_a, etc., and calculated ET as the sum of daytime and nighttime components, thereby enhancing accuracy. The distinction between daytime and nighttime periods was determined based on incoming shortwave radiation (Rs). Specifically, hours with Rs > 10.0 (W m^{-2}) were classified as daytime, while the remainder were defined as nighttime. Although the general framework remained consistent for both periods, distinct parameterization equations were applied for surface conductance (*Gs*) and soil heat flux (*Gsoil*) to account for the differences:

$$G_{S_{night}} = 0.0$$

$$G_{S_{day}} = C_L \times m(Tmin) \times m(VPD) \times r_{corr}$$

$$Gsoil = \begin{cases} 4.73 \times T_i - 20.87 & Tmin_{close} \leq Tann_{avg} < 25^{\circ}C, Tday - Tnight \geq 5^{\circ}C \\ 0.0 \quad Tann_{avg} \geq 25^{\circ}C \; or \; Tann_{avg} < Tmin_{close} & or \; Tday - Tnight < 5^{\circ}C \\ 0.39 * A_i & abs(G) > 0.39 \times abs(A_i) \end{cases}$$

where Gs_{day} and Gs_{night} are daytime and nighttime stomatal conductance, respectively; $Gsoil$ is soil heat flux for bare soil without vegetation, r_{corr} is the total aerodynamic resistance to vapor transport corrected for atmospheric temperature and pressure, $Tann_{avg}$ is annual average daily temperature. C_L, $m(Tmin)$, $m(VPD)$ and $Tmin_{close}$ are empirical parameters based on vegetation cover types. Subsequently, the daily total ET was computed by aggregating the hourly estimates:

$$ET = \frac{\sum_{n=1}^{N} ET_n \times 24}{N}$$

where N is the number of valid hourly observations in a day. More detailed information on the algorithms and the parameters can be found in Mu et al. (2011, 2007).

References:

[1] Mu, Q., Heinsch, F. A., Zhao, M., and Running, S. W.: Development of a global evapotranspiration algorithm based on MODIS and global meteorology data, Remote Sensing of Environment, 111, 519–536, https://doi.org/10.1016/j.rse.2007.04.015, 2007.

[2] Mu, Q., Zhao, M., and Running, S. W.: Improvements to a MODIS global terrestrial evapotranspiration algorithm, Remote Sensing of Environment, 115, 1781–1800, https://doi.org/10.1016/j.rse.2011.02.019, 2011.

2. Lines 163-164: How did you extract the relatively reliable data from global ET products?

3. Lines 168-169: The manuscript mentioned “data from other pixels were excluded to reduce uncertainty”, what is the meaning of other pixels? Do you mean the pixels from other finer products but within the same pixel already settled?

Response: Thank you for your comments. As Question 3 pertains to part of the data extraction process (Step 3) in Question 2, we have combined the responses to Questions 2 and 3. We acknowledge that the original description regarding the global-scale data extraction process was not sufficiently clear. To clarify how reliable data were extracted, we describe the procedure as follows.

Step1: At each pixel for every time point, we calculated the relative errors between estimated ET of three base models and global ET products, obtaining six error values for each base model.

Step2: If a base model yielded relative errors consistently lower than the other two models across all six products, data of this specific pixel and time point was identified as relatively reliable training data, since at this particular spatiotemporal point, the six global products exhibit relative consistency and one base model significantly outperforms the others.

Step3: Data from these specific spatiotemporal points were added into the training data, and the model yielding the lowest relative errors was considered to be the ‘dominant model’ of these specific spatiotemporal points. Data from the remaining spatiotemporal points (referred to as 'data from other pixels' in Question 3) were excluded to reduce uncertainty.

To improve the clarity of the manuscript, we have revised the corresponding description in the Line 163 paragraph as follows:

At the global scale, due to the lack of reliable ET observations, we used six widely used global ET products as references to *extract some relatively reliable data for the training dataset. Specifically,* we calculated the relative errors between estimated ET of base models and global ET products at each pixel for every time point, obtaining six error values for each base model. *If a base model yielded relative errors consistently lower than the other two models across all six products, data of this specific pixel and time point was identified as relatively reliable training data. Data from these specific spatiotemporal points were added into the training data, and the model yielding the lowest relative errors was considered to be the ‘dominant model’ of these specific spatiotemporal points. In contrast, data from the remaining spatiotemporal points were excluded to reduce uncertainty.*

4. Lines 171-173: You get the dominant model for both time and spatial scales. If for the same spatial, the dominant model varies frequently with time, what is your strategy for forecasting? Do you select the dominant model for a specific location with more better-results?

Response: Thank you for your valuable comment. Regarding our forecasting strategy, we implemented a strategy that dynamically selects the dominant model corresponding to each location and time step, rather than fixing a single model for a location. Specifically, even if the dominant model varies frequently with time at the same location, our algorithm utilizes the dominant model identified for each time step for forecasting. This strategy effectively captures the spatiotemporal heterogeneity across different ET models.

In response to your question concerning the dynamic selection strategy, we conducted an additional validation of the 'Static Spatial' Classifier-Guided Ensemble model (i.e., selecting the dominant model for a specific location) using reliable ET datasets at both site and catchment scales, and compared its performance with that of the proposed model.

At the site scale (Fig. R1), the Static Spatial Ensemble model performed reasonably well, with a KGE of 0.91, R^2 of 0.88, MAE of 9.64 mm month^{-1}, and RMSE of 14.31 mm month^{-1}. While it outperformed the individual base models and yielded results comparable to the Attention-based Ensemble, it did not exhibit superior performance compared to the proposed spatiotemporal dynamic Classifier-Guided Ensemble model.

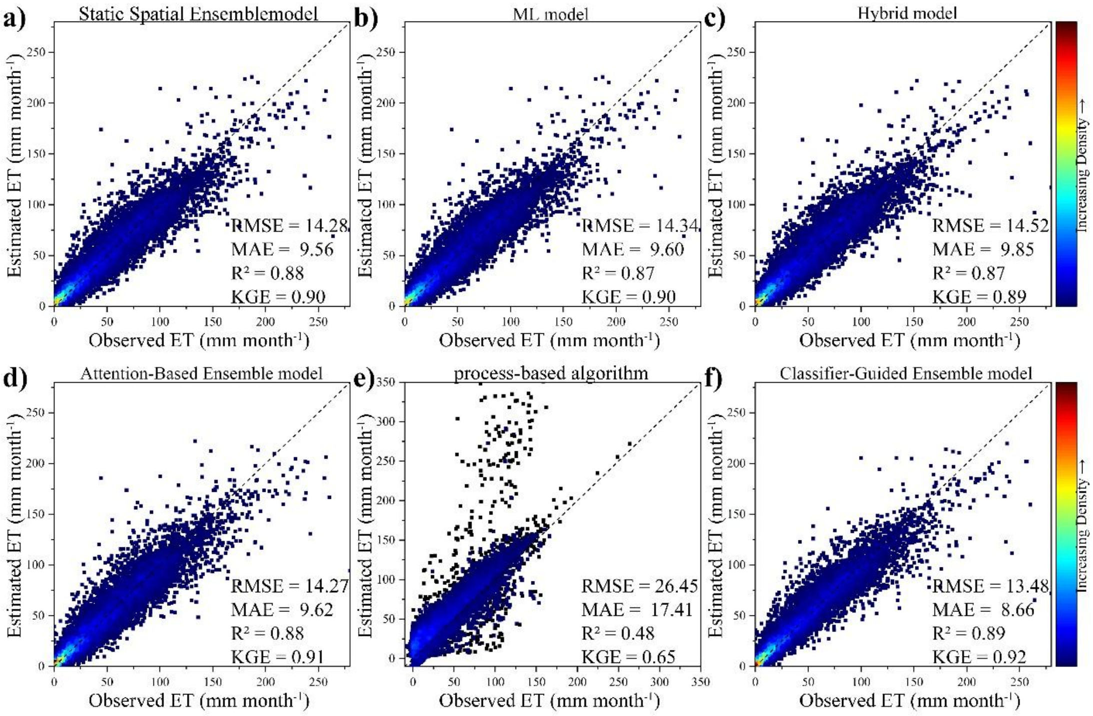
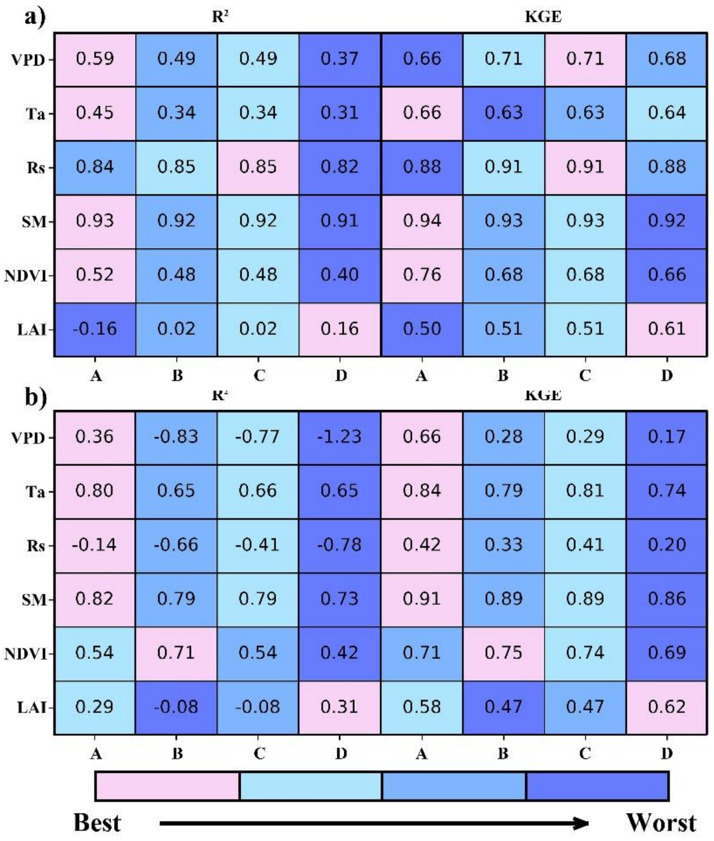


Figure R1. Performance of a) Static Spatial Ensemble model, b) ML model, c) Hybrid model, d) Attention-Based Ensemble model, e) process-based algorithm and f) Classifier-Guided Ensemble model in ten-fold cross-validation.

The underlying reason is that among the base models, the ML model exhibits the highest accuracy at the site scale. Consequently, in the Static Spatial Ensemble model, the ML model is mainly selected as the dominant model. Therefore, similar to the Attention-Based Ensemble model, although process-based and hybrid models are selected at certain sites, the overall performance of the Static Spatial Ensemble model closely mirrors that of the ML model. This similarity persists in the extreme sample validation (Fig. R2): the Static Spatial Ensemble modal performs similarly to the ML model in most extreme samples, showing distinct advantages only under extremely high NDVI conditions. In contrast, the proposed spatiotemporal dynamic Classifier-Guided Ensemble model achieved superior results compared to all base models across the majority of extreme samples.



a)

	R² A	R² B	R² C	R² D	KGE A	KGE B	KGE C	KGE D
VPD	0.59	0.49	0.49	0.37	0.66	0.71	0.71	0.68
Ta	0.45	0.34	0.34	0.31	0.66	0.63	0.63	0.64
Rs	0.84	0.85	0.85	0.82	0.88	0.91	0.91	0.88
SM	0.93	0.92	0.92	0.91	0.94	0.93	0.93	0.92
NDVI	0.52	0.48	0.48	0.40	0.76	0.68	0.68	0.66
LAI	-0.16	0.02	0.02	0.16	0.50	0.51	0.51	0.61

b)

	R² A	R² B	R² C	R² D	KGE A	KGE B	KGE C	KGE D
VPD	0.36	-0.83	-0.77	-1.23	0.66	0.28	0.29	0.17
Ta	0.80	0.65	0.66	0.65	0.84	0.79	0.81	0.74
Rs	-0.14	-0.66	-0.41	-0.78	0.42	0.33	0.41	0.20
SM	0.82	0.79	0.79	0.73	0.91	0.89	0.89	0.86
NDVI	0.54	0.71	0.54	0.42	0.71	0.75	0.74	0.69
LAI	0.29	-0.08	-0.08	0.31	0.58	0.47	0.47	0.62

Figure R2. The comparison of different models (A Classifier-Guided Ensemble model, B Static Spatial Ensemble model, C ML model, D Hybrid model) under extreme conditions in the form of heatmaps. a) and b) represent the extreme samples sorted in ascending order within the 0th – 1st percentiles and 99th - 100th percentiles, respectively.

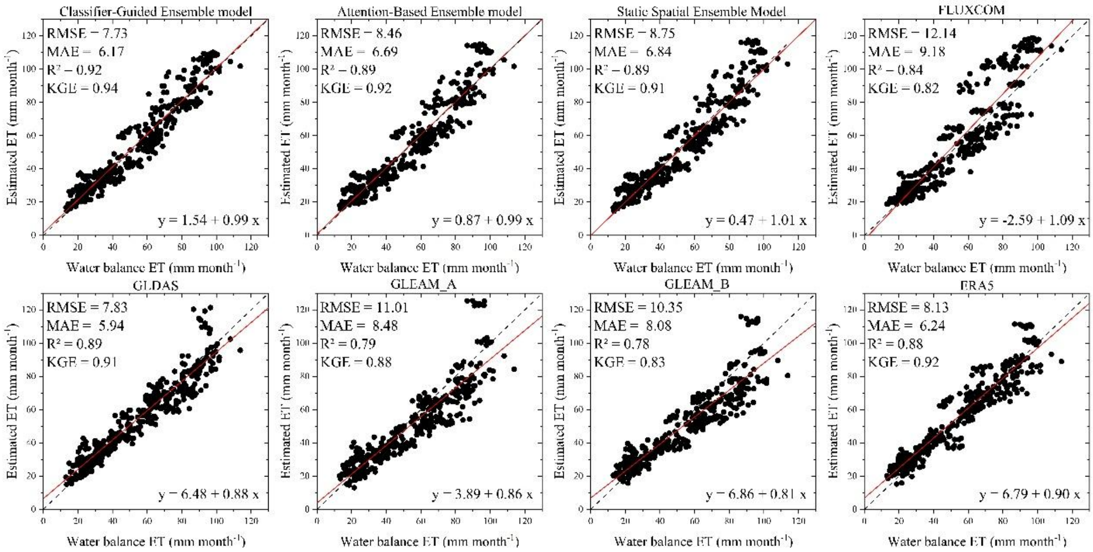


Figure R3. Scatterplot for the relationship between estimated ET and water balance ET (each point represents a catchment over a one-year period).

Similarly, at the basin scale, the Static Spatial Ensemble model did not yield superior performance (Fig. R3). In summary, in addition to the validation across multiple scales and uncertainty analysis (Sections 4 and 5.1), the additional validation further confirms that spatiotemporal dynamic selection is key to optimal accuracy, outperforming the static selection of a dominant model for each location.

5. Line 233: The variables from ERA5-land product were used for training. Were the input variables for the other five models derived from the same data source? If not, were inter-dataset comparisons conducted? If differences among variables are substantial, the training process may introduce model bias.

Response: Thank you for your comments. In this study, all models, including the process-based algorithm, ML-based ET model, Hybrid model, Attention-Based Ensemble Model and Classifier-Guided Ensemble Model, were driven by the same set of input datasets. This consistency ensures that performance differences among our models can be attributed to algorithm structures rather than inconsistent input forcing. To avoid any ambiguity and ensure clarity, we have revised the text in Line 233 as follows:

Input covariates for all models described in Section 2 were derived from the same source to ensure consistency in the training process. Variables P, Rs, SM, ST, Ta, VPD and WS are sourced from the ERA5-land product……

We share your concern regarding uncertainty. Given the current lack of reliable direct global ET observations, we selected six global ET products derived from different data sources and algorithms to conduct a comprehensive validation. As demonstrated in Section 4.3, we verified that our model estimates exhibit high consistency with these reference products in terms of global means, latitudinal profiles, and spatial patterns (Section 4.3.1 and Figure 9). Furthermore, the Bayesian Three-Cornered Hat (TCH) method was employed to quantify the uncertainties associated with both the products and model estimates (Section 4.3.2 and Figure 10). Collectively, these results and validation metrics confirm that these global products serve as reliable benchmarks for our study. We have also revised the text in Line 217 as follows:

Given the current lack of reliable direct global ET observations for global-scale validation, we collected six widely used ET products derived from diverse data sources, forcing data, and calculation methods to conduct a comprehensive validation. (1) ……

6. Figure 5: The colors of Classifier-Guided and Hybrid model are very similar in the axis plots. Using more distinct colors would improve the clarity and comparability of model performance.

Response: Thank you for your suggestion. We apologize for the oversight in the color selection, which affected the clarity of the figure. Figure 5 has been updated with distinct colors as follows:

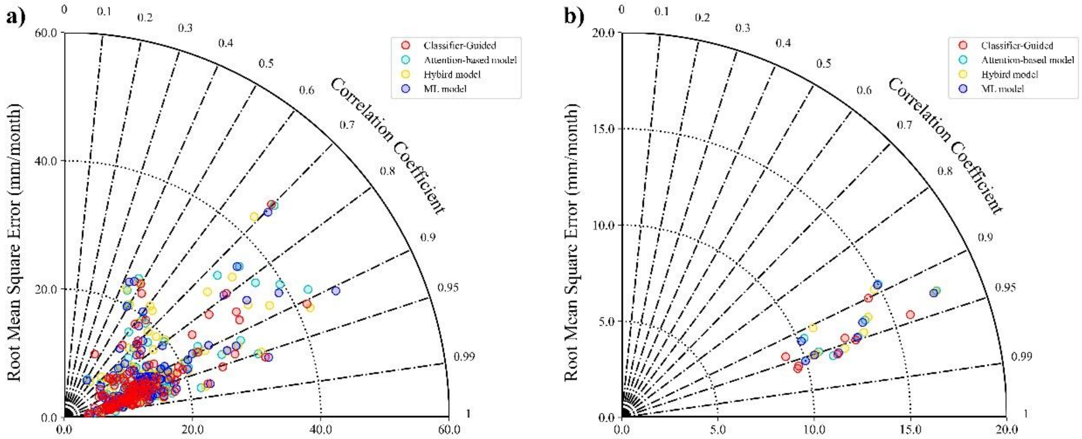


Figure R4. Taylor diagram that compares the performance of the four models with ground observations for a) different sites and b) different land cover types.

7. Lines 557-558: The sentence appears to be incomplete and should be revised for clarity.

Response: Thank you for your careful review. We apologize for the confusion caused by the incomplete sentence structure. We have revised the text in Lines 557-558 as follows:

Also, there is a tendency to select Hybrid model under conditions of low VPD and low SM, while the ML model tends to be selected under high Ta conditions.