Response to Referee #1

The authors conducted a meta-analysis to evaluate the performance of machine learning (ML) algorithms in the estimation of evapotranspiration. I believe this topic is timely and of interest to the HESS community. The motivation of the study, method, and results are clearly outlined, and they reach clear conclusions. Overall, this manuscript is informative and well structured. However, I believe there are several minor aspects which can be improved. Therefore, I support its publication in HESS with minor revisions.

Response: We would like to thank the reviewer for the positive comments and the time invested to review our manuscript. The revised manuscript will follow the reviewer's recommendations.

1) L34 "ET is the most important indicator of the water cycle": ET is not an indicator. It is a water balance component. Also, it may be not the most important component. I suggest writing "ET is one of the most important components of the water cycle ~"

Response: Thank you for the insightful comments. It will be revised as 'ET is one of the most important components of the water cycle'.

2) L51-53: add examples and references to support the argument.

Response: Two references will be added: 'For remote sensing-based physical models and process-based land surface models, some physical processes have not been well characterized due to the lack of understanding of the detailed mechanisms influencing ET under different environmental conditions. For example, the inaccurate representation and estimation of stomatal conductance (Li et al., 2019) and the linearization (McColl, 2020) of the Clausius-Clapeyron relation in the Penman-Monteith equation may introduce both empirical and conceptual errors into estimates of ET.'

3) L82: define NDVI, EVI and LAI.

Response: It will be defined as Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), and Leaf area index (LAI).

4) L83: define GPP

Response: It will be defined as 'Gross Primary Productivity.

5) L153-155: I agree with the authors' point, but RMSE is still an important measure of the model performance. I think there is a way to normalize the RMSE when the magnitude or standard deviation of water flux are available. If possible, I recommend analyzing RMSE as well.

Response: Thank you for the insightful comments. The RMSE depends on the magnitude of the ET value of the training data. For example, due to the difference in the range of ET values, models developed from flux stations in dry grasslands will typically have lower RMSE than models developed by flux stations based on forests in wet areas. Therefore, RMSE may not be a good metric for cross-study comparisons. We will clarify this issue in the revised manuscript. Since we do not have the raw data of these studies, it is difficult to unify the differences in RMSE across data sets in a normalized way.

Mean Absolute Percentage Error (MAPE) can be useful but not commonly used or reported as R-squared in such studies.

6) L225-229 and Figure 5 and Figure7: I think the authors should discuss variables which decrease the performance of the ML models (NDVI etc.). To do this, the authors may need to refer to Figure 7. Therefore, I suggest reordering Figures (i.e., 7 ->6 and 6->7). Figure 7 implies performance decreases due to NDVI (and other variables) may be spurious. In order to overcome such limitations, I suggest performing additional analysis by grouping ML models which use Rn/Rs and Ta and then generating Figure 5.
Response: Thank you for the insightful comments. This is a good suggestion. We will consider adjusting the order of the figures based on your comments and will perform additional analysis by grouping ML models which use Rn/Rs and Ta as you suggested.

7) Figure5: difficult to compare variables. I think visualization can be improved by grouping variables which improve performance or not.
Response: We will consider adjusting the order of the figures based on your comments, and will perform additional analysis by grouping variables as you suggested (also based on findings in Fig. 7).

8) L261-263: I cannot agree. Data-driven approach and process-based approach are complementary. This should be revised.
Response: We will modify the description here. Indeed data-driven and process-based approaches are complementary and both are rapidly developing and therefore of equal importance in the future direction of ET estimation.

9) L336-338: As the authors briefly mentioned here, eddy covariance observations are subject to random, gap-filling, and systematic (energy balance closure) uncertainty. There are several ways to address this uncertainty. For example, some studies may use a gap-filled dataset but some studies may choose observation only. Also, the energy balance closure problem can be addressed differently (uncorrected, Bowen-ratio corrected, and use of energy balance residual). Depending on this choice, the performance of ML algorithms may vary significantly (particularly energy closure problem is important). Although the authors mentioned observational uncertainty as a limitation of this research in L336-338, I believe this brief mention is not enough. If you can extract this information from the literature, I suggest performing an additional analysis (e.g., performance comparison for energy balance corrected vs uncorrected). If it is indeed difficult to extract the information from the literature, this topic should be discussed more thoroughly at least.
Response: We will elaborate on the discussion section on this issue. Indeed uncertainties in the observations (including those in Gap-filling) may affect model accuracy. The energy closure problem does also confuse researchers in this field which may lead to the underestimation of ET values, although some datasets (e.g., FLUXNET) have provided observations of latent heat after bias correction in energy closure.

When the problem of energy closure is not negligible, the use of energy balance uncorrected data may affect the model accuracy. We will discuss this issue further based on previous studies (combined with the potential severity of the bias in ET observations caused by the energy closure problem in various environmental conditions).

References

Li, X., Kang, S., Niu, J., Huo, Z., and Liu, J.: Improving the representation of stomatal responses to CO2 within the Penman–Monteith model to better estimate evapotranspiration responses to climate change, Journal of Hydrology, 572, 692–705, https://doi.org/10.1016/j.jhydrol.2019.03.029, 2019.

McColl, K. A.: Practical and Theoretical Benefits of an Alternative to the Penman-Monteith Evapotranspiration Equation, 56, e2020WR027106, https://doi.org/10.1029/2020WR027106, 2020.