Response to Referee #2

In this study, Shi et al., presented a meta-analysis of the performance of machine learning (ML) algorithms in the estimation of evapotranspiration. While this manuscript is interesting and within the scope of HESS, I have a few major concerns.

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.

Most importantly, while this is a meta-analysis, the authors were comparing results from different publications, in which different data sets and sites may have been used. That being said, some of the results are not directly comparable. For example, Zeng et al. 2020, may have selected a few sites that are much more difficult to predict; and can not be compared with the results presented in another publication. Also, some sites may use in-situ estimates of LAI and VIs, while others use LANDSAT or even MODIS LAI and VIs. In order to make their results publishable, they need to find a way to harmonize the data sets used in all studies. Or, they need to justify that they have an inclusion criteria when selecting all publications (instead of just stating we searched on Scopus). In addition, I am not sure whether the number of models they chose can well support their comparison of so many features.

Response: Thank you for the insightful comments. Some studies have indeed used sites that are difficult to predict. Usually, with meta-analysis, we only get comprehensive findings, and it is difficult to improve the understanding of extreme and exceptional cases (because the mean or median of statistical results is what we used in the formal assessment). The inclusion of extreme cases (such as the very unpredictable sites you mentioned) may negatively affect the evaluation results, but this negative effect may be limited if they only share a low proportion of the samples.

In addition, there are comparisons of studies using the same data (but different algorithms) (Fig. 3b) in this study. The difference in the data between studies is constrained (keeping other features the same but only the algorithms different): Fig 3a included various conditions across studies, i.e., what the reviewers raised; Fig 3b is the result of a comparison of model cases based on the same data and different machine learning algorithms, and is a correction and a more objective characterization of the issue with Fig 3a.

Few studies have used in-situ measured LAIs and VIs for modeling, as this is not helpful for the large-scale, long-time series predictions compared to remote sensing-based LAIs and VIs. Regarding these worries, we will clarify the details of these inclusion criteria which were used for the screening of the article in the revised manuscript.

Although multiple features were evaluated in this study, there are only a few features that predominantly affect the accuracy of the model. Some features may be insignificant (only weakly influencing) and we will consider deleting these features to highlight the analysis of

the major influencing features. In addition, we have included as large a sample as possible to support our findings, and our findings for the meta-analysis of ET predictions are likely to be more robust as further such studies are added in the future.

Also, the authors have another paper looking at similar topics (even with some similar pictures and texts) in discussion on Biogeoscience. As an example, in this paper:
Line 114-117: And in machine learning, in general, modeling with unbalanced data (with significant differences in the distribution between the training validation sets) may result in lower model accuracy.
And in the BG paper:
Line 91-94: Modeling with unbalanced data (where the difference between the distribution of the training and validation sets is significant even if selected at random) may result in lower model accuracy.
The only differences between the two papers is that the BG paper focused on NEE, while this paper looked into ET. I am not sure whether it is acceptable to publish two somewhat similar papers in two different EGU journals.
Response: Thank you for the insightful comments. Although the framework/methods of these two manuscripts are similar, the topics are different. One is NEE (which is a carbon cycle-related topic) and the other is ET (which is a water cycle-related topic). NEE and ET are not directly correlated and therefore need to be studied separately. The NEE and ET prediction use different explanatory variables and analysis/discussions of their mechanisms are also different. The potential readers of these two manuscripts are also different. We will carefully check for possible duplicate text.

At the same time, overall, the writing of the manuscript is good. But I do find it difficult to follow from time to time. For example, the authors used many abbreviations without defining them (EVI, GPP and NDVI), and some of them may not be very familiar with all the readers.
Response: Abbreviations will be defined such as Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), and Leaf area index (LAI).

Minor comments:

Line 34: I suggest that the authors refrain from statements like this, precipitation and runoff are at least equally important;
Response: Thank you for the insightful comments. It will be corrected as 'ET is one of the most important components of the water cycle in terrestrial ecosystems.'

Line 52: detailed?
Response: Two references will be added (for the detailed limitations in physics-based methods): '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.'

Line 155: I still do not understand why RMSEs are not used.
Response: 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.

L186: outperformed whom? I believe that similar issues can be found in other places of the manuscript.
Response: The 'outperform' here refers to the higher accuracy of SVM and RF compared to other algorithms in Fig. 3a. We will further check other such descriptions in this manuscript.

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.