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'.

Action: revised as 'Evapotranspiration (ET) is one of the most important components of the water cycle in terrestrial ecosystems.'

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.'

Action: elaborated as '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).

Action: Defined.

4) L83: define GPP

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

Action: Defined.

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 humid regions. 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.

Action: clarified as 'Although RMSE is also often used for model accuracy assessment, its dependence on the magnitude of water flux values makes it difficult to use for fair comparisons between studies. 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 humid regions. Therefore, RMSE may not be a good metric for cross-study comparisons.'

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.

Action: We placed Figure 7 in the supplementary material as Figure S2 and replaced Figure 5 with an figure of evaluation of the combination of predictor variables. Figure 5 was placed in the additional material as Figure S1. In the revised manuscript, we assessed the impact of predictor variables on model accuracy at two levels: (1) the correspondence between the use or non-use of individual predictor variables and model accuracy, and (2) after analyzing the dependence of the use of predictor variables, we analyze the impact of combinations of predictor variables (distinguishing mainly between meteorological predictor combinations and remote sensing-based vegetation-related predictor combinations).

This paragraph was elaborated as:

'On one hand, from the perspective of the effect of individual predictors, the use of Rn/Rs, Prec, Ta, and FAPAR improved the accuracy of the model (Fig. S1). This pattern partially changed in the different PFTs. In the forest sites, the accuracy of the models with Rn/Rs and Ta used was higher than that of the models with Rn/Rs and Ta not used. For the

grassland sites, the use of Ws, FAPAR, Prec, and Rn/Rs improved the model accuracy. For the cropland sites, Ta and FAPAR were more important for improving the model accuracy.

On the other hand, the evaluation of the effect of individual predictors on model accuracy is not necessarily reliable because some predictor variables are used together (e.g., the high model accuracy corresponding to a particular variable may be due to the fact that it is often used together with another variable that really plays the dominant role in improving accuracy). Therefore, we tested for independence between the use of variables and assessed the effect of the combination of variables on model accuracy. We calculated the correlation matrix (Fig. S2) between the use of various predictors (not used is set as 0 and used is set as 1). We found there was dependence between the use of some predictors, the use of NDVI/EVI, LAI, and SM was significantly negatively correlated with the use of Rn/Rs and Ta (Fig. S2). It indicated that many of the models that used Rn/Rs and Ta did not use NDVI/EVI, LAI, and SM, and the models that used NDVI/EVI, LAI, and SM also happened to not use Rn/Rs and Ta. Given this dependence between the use of predictors, we evaluated the effect of the combination of variables on the model accuracy (Fig. 5). In Fig. 5, the three variable combinations on the left side are mainly meteorological variables while the three variable combinations on the right side are mainly vegetation-related variables based on remote sensing (e.g., NDVI, EVI, LAI, LSWI). We found that, overall, the accuracy of the models using only meteorological variable combinations was higher than that of the models using only remote sensing-based vegetation-related variables. It demonstrated the importance of using meteorological variables in machine learning-based ET prediction (probably especially for models with small time scales such as hourly scale, daily scale). For example, in the forest type, the combination of Ta and Rn/Rs is very effective compared to using only remote sensing-based vegetation index variable combinations. The combination of Ta and Rn/Rs is also effective in the grassland and cropland types. The combination of Ws and Rn/Rs played an important role in the grassland type for improving model accuracy. Despite this, it does not negate the positive role of remote sensing-based vegetation-related variables in ET prediction. This effectiveness can be dependent on the time scale of the model as well as the PFTs. In models with large time scales (monthly scale, seasonal scale) and PFTs in which ET is sensitive to vegetation dynamics, remote sensing-based vegetation-related variables may also be of high importance.’

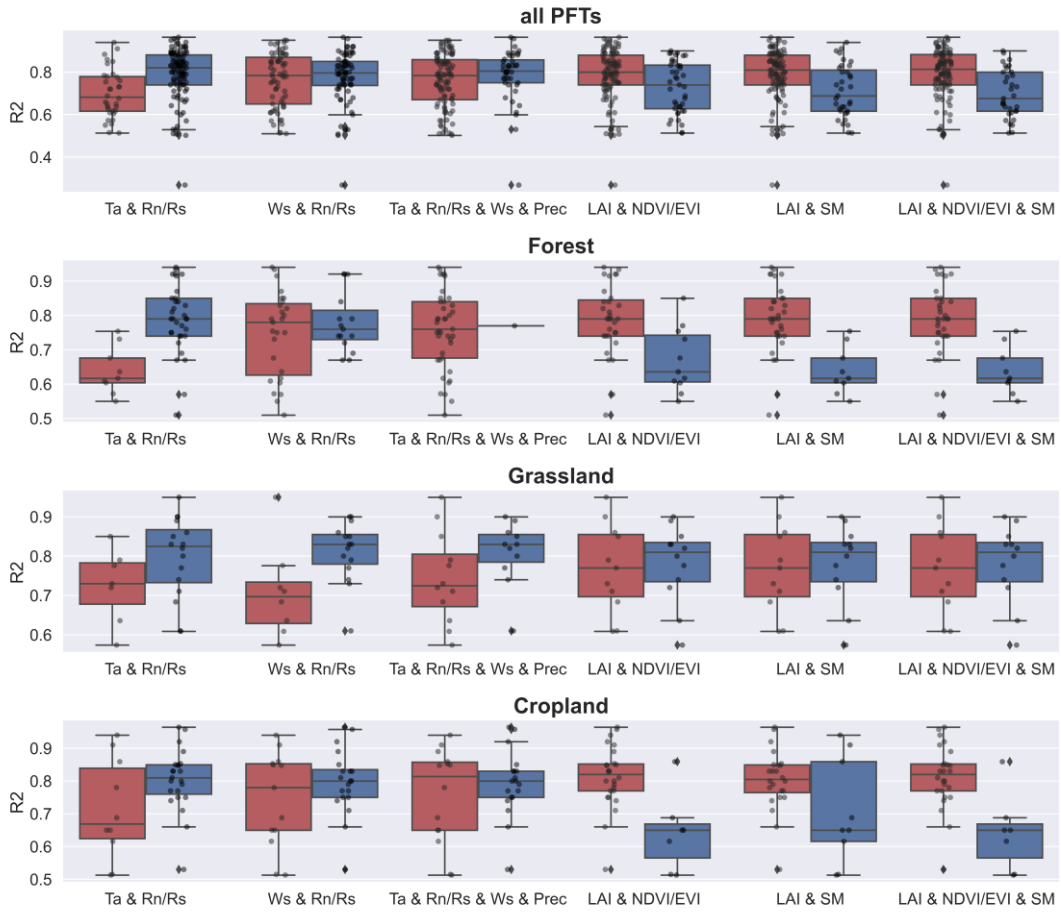

Figure 5. Effects of combinations of predictor variables on model accuracy in various PFTs (all data, forest, grassland, and cropland). Dark blue boxes indicate that the predictors were together used in the model (e.g., for 'Ta & Rn/Rs', the dark blue box represents Ta and Rn/Rs were together used in the model), while dark red boxes indicate the other conditions (i.e., the combination was not used). Predictors: precipitation (Prec), soil moisture/remote sensing-based land surface water index (SM), net radiation/solar radiation (Rn/Rs), enhanced vegetation index (EVI), air temperature (Ta), leaf area index (LAI), Normalized Difference Vegetation Index/Enhanced Vegetation Index (NDVI/EVI).

We also added a paragraph in the discussion section on the use of predictor combinations:
'Biophysical and meteorological variables are considered both important in ET simulations. This study found that models using a combination of meteorological variables had higher accuracy than models using only remotely sensed vegetation dynamic information. However, due to the high proportion of models with small temporal scales (e.g., half-hourly scale, hourly scale, and daily scale) in this study, this advantage of the combination of meteorological variables may be more suitable to small temporal scales. A possible explanation is that vegetation-related variables such as NDVI and LAI at the daily scale, 8-day scale, and 16-day scale have limited explanatory ability for hourly or daily-scale variability in ET. At small temporal scales, the use of combinations of meteorological variables can capture moisture and energy conditions that control the rapid fluctuations of ET and thus has a dominant role in hourly or daily-scale ET prediction. This also corroborates with the high accuracy of some physic-based ET estimation models (Rigden and

Salvucci, 2015) that use only meteorological variables and not vegetation-related variables such NDVI (only an estimate of vegetation height derived from land cover maps is used to represent vegetation conditions (Rigden and Salvucci, 2015)).'

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).
Action: In the revised manuscript, we evaluated the effect of the combination of predictor variables (please refer to the above response/action in the last comment).

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.
Action: revised as 'With the accumulation of in situ EC observations around the world, the study of ET simulations based on data-driven approaches has received more attention from researchers in the last decade. Many studies have combined EC observations, various predictors, and machine learning algorithms to improve the prediction accuracy of site-scale water fluxes.'

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).

Action: the discussion section 4.2.3 was elaborated:

(a) Uncertainties caused by data quality control (e.g. gap-filling (Hui et al., 2004)) are difficult to assess effectively. Gap-filling is a commonly used technique to fill in low-quality data in flux observations (Chen et al., 2012; Hui et al., 2004). However, the impact of this practice on machine learning-based ET prediction models is unclear, due to the difficulty of directly assessing how this technique is performed in various studies by this meta-analysis. Typically, models with small temporal scales (e.g., hourly scale, daily scale) can exclude low-quality observations and use only high-quality data. However, for models with large time scales (e.g., monthly scales), gap-filling (e.g., based on meteorological data) may be unavoidable. This may lead to decrease in training data purity and introduce uncertainty in the subsequent prediction model development.

(b) Systematic uncertainties caused by the energy balance closure (EBC) issue in eddy-covariance flux measurements are also difficult to assess by this meta-analysis. EBC is a common problem (Eshonkulov et al., 2019) in eddy-covariance flux observations. For that reason, the latent heat flux measured potentially underestimates ET. Some prediction models corrected EBC (e.g., using Bowen ratio preserving (Mauder et al., 2013, 2018) and energy balance residuals (Charuchittipan et al., 2014; Mauder et al., 2018)) in the processing of training data, but some did not. How this will affect the accuracy of the prediction model is not clear due to multiple factors that need to be evaluated that influence EBC (Foken, 2008), including measurement errors of the energy balance components, incorrect sensor configurations, influences of heterogeneous canopy height, unconsidered energy storage terms in the soil-plant-atmosphere system, inadequate time averaging intervals, and long-wave eddies (Jacobs et al., 2008; Foken, 2008). To reduce this uncertainty, more attention to flux site characteristics (Eshonkulov et al., 2019) related to PFT, topography, flux footprint area, etc., to select the appropriate correction method is necessary for future studies.

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.

Actions:

Action 1:

Further response:

The purpose of meta-analysis is to combine the heterogeneity of studies to obtain comprehensive findings. If we filter all articles to use the criteria 'using the same data and sites', then few articles can be included in the meta-analysis and this analysis will be difficult to implement.

The paper by Zeng et al. uses global-scale data from FLUXNET (which simulates carbon fluxes rather than water flux) with high variations in site conditions, some sites of which may indeed be much more difficult to predict. Such a global scale using FLUXNET data is the largest in such studies and belongs to the outlier/extreme cases. This meta-analysis gave a general/average reference for such studies, and for very large scale, or otherwise specific studies, researchers should still focus on the specificity of their models.

On the feasibility of the methodology to assess the impacts of various model features on model accuracy by meta-analysis, there are several published articles (listed below) using a similar methodology in the cross-study comparisons (comparison of models developed in various studies despite the different data and features used).

- Khatami, R., Mountrakis, G., and Stehman, S. V.: A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research, Remote Sensing of Environment, 177, 89–100, https://doi.org/10.1016/j.rse.2016.02.028, 2016.
- Zolkos, S. G., Goetz, S. J., and Dubayah, R.: A meta-analysis of terrestrial aboveground biomass estimation using lidar remote sensing, Remote Sensing of Environment, 128, 289–298, https://doi.org/10.1016/j.rse.2012.10.017, 2013.

Therefore, if the objective is to obtain a comprehensive understanding and general guidelines, we think the meta-analysis methodology of this study can be feasible (although some special cases should also be attended to).

We hope you can agree on the value of this meta-analysis, although some of such studies can vary widely and thus were not very comparable.

Action 2:

We clarified the detailed inclusion criteria when selecting all publications and the way to harmonize the data sets used in all studies:

Revised as:

'We applied a general query (on December 1st, 2021) on title, abstract, and keywords to include articles with the "OR" operator applied among expressions (Table 1) in the Scopus database. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Moher et al., 2009) are followed when filtering the papers. We first excluded articles that obviously did not fit the topic of this study based on the abstract, and then performed the article screening with the full-text reading.

The inclusion of articles follows the following criteria:

a) Articles were filtered for those with water fluxes (or latent heat) simulated.
b) The water flux or latent heat observations used in the prediction models should be from the eddy-covariance flux measurements.
c) Articles focusing only on gap-filling techniques (i.e., the objective was not simulation and extrapolation of water fluxes using machine learning) were excluded.
d) Only articles that used multivariate regression (with the number of covariates greater than or equal to 3) were included.
e) The determination coefficient (R-squared) of the validation step should be reported as the metric of model performance (Shi et al., 2021; Tramontana et al., 2016; Zeng et al., 2020) in the articles.
f) The articles should be published in English-language journals.'

Besides, a column 'Harmonization' is added to Table 2 to describe how to harmonize the features used in all studies into specific categories.

Table 2. Description of information extracted from the included papers.

Field	**Definition & Categories adopted**	**Harmonization**
Climate	Climate zones of the study location derived from the Köppen climate classification (Peel et al., 2007)	
Plant functional type (PFT)	PFT of the flux sites: 1-forest, 2-grassland, 3-cropland, 4-wetland, 5-shrubland, 6-savannah, and multi-PFTs	The categorization is based on the descriptions in the article. For example, cropland for various crops is classified as 'cropland', and both woody savannah and savannah are classified as 'savannah'.
Location	More precise location (with the latitude and longitude of the center of the studied sites): latitude, longitude	
Algorithms	Random Forests (RF), Multiple Linear Regressions (MLR), Artificial Neural Networks (ANN), Support Vector Machines	Various model algorithms with parameter optimization or other improvements are categorized as their algorithm family. For

	(SVM), Cubist, model tree ensembles (MTE), K-nearest neighbors (KNN), long short-term memory (LSTM), gradient boosting regression tree (GBRT), extra tree regressor (ETR), Gaussian process regression (GPR), Bayesian model averaging (BMA), extreme learning machine (ELM), and deep belief network (DBN)	example, various improved models of RF algorithms are classified as RF, rather than as another algorithm family.
Sites number	Number of the flux sites used	
Spatial scale	Area representatively covered by the flux sites: local (less than 100 x 100 km), regional, global (continent-scale and global scale)	The spatial scale is roughly categorized based on the area covered by the site. The model is classified as 'global' only when the spatial extent reaches the continental scale.
Temporal scale	The temporal scale of the model: half-hourly, hourly, daily, 4-daily, 8-daily, monthly, seasonally (i.e., 0.02, 0.04, 1, 4, 8, 30, 90 days)	Models with a temporal scale greater than one month and less than one year are classified as seasonal scale models.
Year span	The span of years of the flux data used	Year span is calculated as the span from the earliest to the latest year of available flux data.
Site year	Describe the volume of total flux data with the number of sites and years aggregated.	
Cross-validation	Describe the chosen method of cross-validation: Spatial (e.g., 'leave one site out'), temporal (e.g., 'leave one year out'), random (e.g., 'k-fold')	
Training/validation	Describe the ratio of the data volume in the training and validation sets.	In spatial validation, this ratio is represented by the ratio of the number of sites used for training to the number of sites used for validation. In temporal validation, this is represented by the ratio of the span of time periods used for training to the span of time periods used for validation.
Satellite images	Describe the source of satellite images used to derive NDVI, EVI, LAI, LST, etc: Landsat, MODIS, AVHRR	

Biophysical predictors	LAI, NDVI/EVI, the fraction of absorbed photosynthetically active radiation/photosynthetically active radiation (FAPAR/PAR), leaf area index (LAI), Carbon fluxes (CF) including NEE/GPP, etc.	The predictor variables of different measurement methods are categorized according to their definitions. For example, both using the NDVI calculated based on satellite remote sensing bands and in situ measurements were classified as the use of 'NDVI'.
Meteorological variables	precipitation (Prec), net radiation/solar radiation (Rn/Rs), air temperature (Ta), vapour-pressure deficit (VPD), relative humidity (RH) , etc.	The way meteorological data are measured is not differentiated. For example, both using Ta from reanalysis data and Ta measured at flux sites were classified as the use of Ta.
Ancillary data	Describe the ancillary variables used: soil texture, terrain (DEM), soil moisture/land surface water index (SM/LSWI), etc.	Both the use of in situ measured soil moisture and the use of remote sensing-based LSWI was classified as using surface moisture-related indicators SM/LSWI.
Accuracy measure	Accuracy measure used to assess the model performance: R-squared (in the validation phase)	

Action 3:

For the question of whether the sample size of the model supports our findings (evaluation of multiple features), a further analysis was performed: the linear correlations of the model features and their significance were calculated for the R-squared and quantitative. In the context of a limited sample size, tests of statistical significance may provide the readers with some information about the reliability and significance of our findings. The results showed that some features showed positive but not significant correlation with the R-squared. At the same time, some characteristics have a significant positive correlation with the R-squared (e.g., the use of variable combinations). We analyzed these issues specifically in Section 3.3.5.

It is elaborated as:

3.3.5 Linear correlation of quantitative features and R-squared

We also analyzed the linear correlation (Fig. 7) between multiple quantitative features and the R-squared. We found that the magnitude of the linear correlation coefficients between the use of predictor combinations and the R-squared was higher than other features. The use of the predictor combination 'Ta and Rn/Rs' significantly improved the model accuracy. 'Temporal scale', 'time span', 'training/validation ratio', and 'number of sites' showed weak positive correlations with R-squared (not significant, p-value > 0.1). The positive correlation between 'temporal scale' and

R-squared is higher among these features, although not significant. It should also be paid more attention to in future studies. The feature 'training/validation ratio' and 'time span' are also positively correlated (although not significantly) with the R-squared, suggesting the importance of the volume of data in the training set in a data-driven machine learning model. Larger 'training/validation ratio' and 'time span' may correspond to greater proportional coverage of the scenarios/conditions in the training set over the validation set, and thus correspond to higher accuracy.

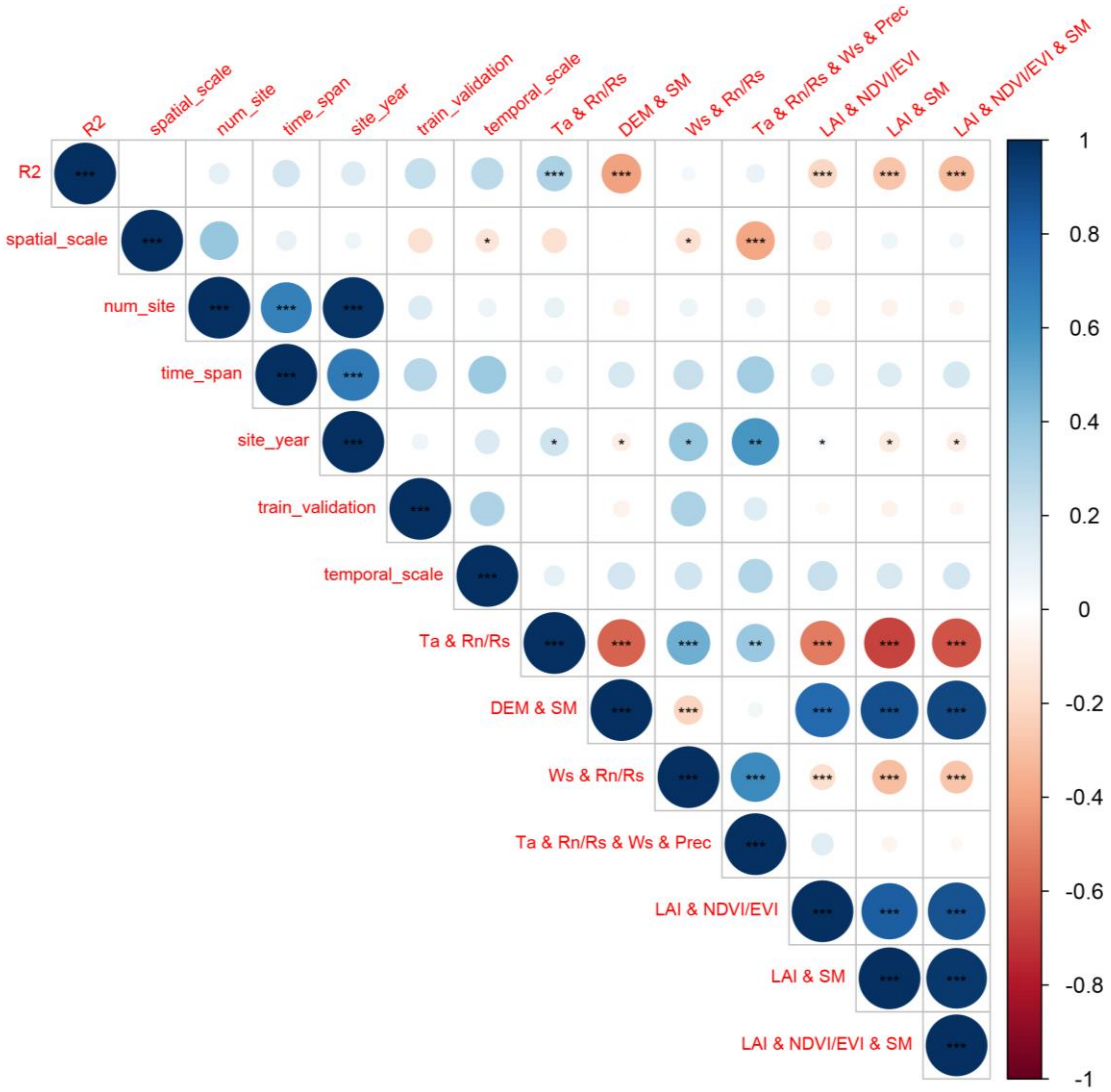

Figure 7. Evaluation of linear correlations between multiple features and the R-squared records with the statistical significance test. For the feature 'spatial scale', the 'local' scale was set to 1, the 'regional' scale was set to 2, and the 'global' scale was set to 3 in the analysis of linear correlation. For the use of various predictor combinations with '&', the value for 'together used' is set as 1 and other conditions are set as 0 (e.g., for the feature 'Ta & Rn/Rs & Ws & Prec', if Ta, Rn/Rs, Ws, and Prec were used together in the model, the value is set as 1). Significance: the p-value < 0.01 (***), 0.05 (**), and 0.1 (*).

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.

Action: We have checked for possible duplicate text in these two manuscripts.

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).
Action: These abbreviations were defined.

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.'
Action: Corrected as 'Evapotranspiration (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.'

Action: elaborated as '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.
Action: clarified as 'Although RMSE is also often used for model accuracy assessment, its dependence on the magnitude of water flux values makes it difficult to use for fair comparisons between studies. 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 humid regions. Therefore, RMSE may not be a good metric for cross-study comparisons.'

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.
Action: revised as 'SVM and RF outperformed (Fig. 3a) across studies (better than other algorithms with sufficient sample size in Fig. 3a such as ANN).'

'With sufficient sample size' were also added to other such descriptions.

References

Charuchittipan, D., Babel, W., Mauder, M., Leps, J.-P., and Foken, T.: Extension of the Averaging Time in Eddy-Covariance Measurements and Its Effect on the Energy Balance Closure, Boundary-Layer Meteorol, 152, 303–327, https://doi.org/10.1007/s10546-014-9922-6, 2014.

Chen, Y.-Y., Chu, C.-R., and Li, M.-H.: A gap-filling model for eddy covariance latent heat flux: Estimating evapotranspiration of a subtropical seasonal evergreen broad-leaved forest as an example, 468–469, 101–110, https://doi.org/10.1016/j.jhydrol.2012.08.026, 2012.

Eshonkulov, R., Poyda, A., Ingwersen, J., Wizemann, H.-D., Weber, T. K. D., Kremer, P., Högy, P., Pulatov, A., and Streck, T.: Evaluating multi-year, multi-site data on the energy balance closure of eddy-covariance flux measurements at cropland sites in southwestern Germany, 16, 521–540, https://doi.org/10.5194/bg-16-521-2019, 2019.

Foken, T.: The energy balance closure problem: An overview, Ecological Applications, 18, 1351–1367, 2008.

Hui, D., Wan, S., Su, B., Katul, G., Monson, R., and Luo, Y.: Gap-filling missing data in eddy covariance measurements using multiple imputation (MI) for annual estimations, 121, 93–111, https://doi.org/10.1016/S0168-1923(03)00158-8, 2004.

Jacobs, A. F. G., Heusinkveld, B. G., and Holtslag, A. A. M.: Towards Closing the Surface Energy Budget of a Mid-latitude Grassland, Boundary-Layer Meteorol, 126, 125–136, https://doi.org/10.1007/s10546-007-9209-2, 2008.

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.

Mauder, M., Cuntz, M., Drüe, C., Graf, A., Rebmann, C., Schmid, H. P., Schmidt, M., and Steinbrecher, R.: A strategy for quality and uncertainty assessment of long-term eddy-covariance measurements, Agricultural and Forest Meteorology, 169, 122–135, https://doi.org/10.1016/j.agrformet.2012.09.006, 2013.

Mauder, M., Genzel, S., Fu, J., Kiese, R., Soltani, M., Steinbrecher, R., Zeeman, M., Banerjee, T., De Roo, F., and Kunstmann, H.: Evaluation of energy balance closure adjustment methods by independent evapotranspiration estimates from lysimeters and hydrological simulations, 32, 39–50, https://doi.org/10.1002/hyp.11397, 2018.

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.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., and Prisma Group: Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement, PLoS medicine, 6, e1000097, 2009.

Peel, M. C., Finlayson, B. L., and McMahon, T. A.: Updated world map of the Köppen-Geiger climate classification, 11, 1633–1644, https://doi.org/10.5194/hess-11-1633-2007, 2007.

Shi, H., Hellwich, O., Luo, G., Chen, C., He, H., Ochege, F. U., Van de Voorde, T., Kurban, A., and de Maeyer, P.: A global meta-analysis of soil salinity prediction integrating satellite remote sensing, soil sampling, and machine learning, 1–15, https://doi.org/10.1109/TGRS.2021.3109819, 2021.

Tramontana, G., Jung, M., Schwalm, C. R., Ichii, K., Camps-Valls, G., Ráduly, B., Reichstein, M., Arain, M. A., Cescatti, A., Kiely, G., Merbold, L., Serrano-Ortiz, P., Sickert, S., Wolf, S., and Papale, D.: Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms, Biogeosciences, 13, 4291–4313, https://doi.org/10.5194/bg-13-4291-2016, 2016.

Zeng, J., Matsunaga, T., Tan, Z.-H., Saigusa, N., Shirai, T., Tang, Y., Peng, S., and Fukuda, Y.: Global terrestrial carbon fluxes of 1999–2019 estimated by upscaling eddy covariance data with a random forest, 7, https://doi.org/10.1038/s41597-020-00653-5, 2020.