

Dear Editor and Reviewers,

We sincerely thank the Editor and the Reviewers for their careful evaluation of our manuscript “Technical note: How well do evapotranspiration partitioning approaches perform in moss-covered wetlands?” submitted to *EGU Biogeosciences*, and for the constructive comments and suggestions. We greatly appreciate the time and effort devoted to reviewing our work. The comments have significantly improved the quality, rigor, and clarity of the manuscript.

The reviewers raised several important suggestions, including the incorporation of a machine learning-based ET partitioning approach, a more rigorous evaluation of method performance, an investigation of the effects of flooding on the evaluated methods, and more detailed descriptions of the study sites and the independent measurements used for model validation. Following the reviewers' suggestions, we have thoroughly revised the manuscript. While these revisions have led to several new findings and strengthened our interpretation of the results, our main conclusion remains unchanged: no single ET partitioning approach consistently outperforms the others across the studied wetlands, and applying multiple approaches concurrently, whenever possible, is recommended to identify sources of uncertainty and potential errors.

The major revisions are summarized below.

First, following the reviewers' recommendation, we incorporated the machine learning framework developed by Stapleton et al. (2022). Because this framework is designed to identify the optimal machine learning algorithm and predictor feature set for each site, we trained and evaluated the complete framework but only included the best-performing machine learning model for each study site in the comparison with the existing ET partitioning approaches. This allows a direct comparison between the strongest machine learning implementation and the other ET partitioning methods.

Second, we expanded model performance evaluation using two complementary metrics. The coefficient of determination (R^2) is used to quantify how well the predicted T:ET follows the temporal variability of the reference data, whereas the Nash-Sutcliffe model efficiency coefficient (NSE) evaluates whether model predictions perform better than simply using the observed mean as the predictor. In addition to the independent measurement-based T:ET constructed from micro-lysimeter, sap flow, chamber, and eddy covariance measurements, we introduced four ensemble datasets. These include the ensemble mean of all available ET partitioning approaches (all-method

ensemble), the ensemble mean of the high-frequency EC-based methods (HF ensemble), the ensemble mean of the ecosystem carbon-water coupling methods (CW ensemble), and the ensemble mean of the machine learning approaches (ML ensemble). The ensemble mean is calculated as the average prediction across the available methods within each group. It is expected to reduce the influence of method-specific errors and provides a more robust reference than any single approach. Individual methods and group ensembles were therefore evaluated against both the measurement-based T:ET and the all-method ensemble.

Third, we added a Spearman correlation analysis to evaluate the consistency among individual methods and their corresponding group ensembles. This analysis provides additional insight into the extent to which different approaches produce consistent or contrasting estimates of T:ET.

Fourth, to investigate whether flooding deteriorates ET partitioning performance, we compared the R^2 and NSE values obtained using the complete Burstall dataset with those obtained after excluding the flooding period. We only analysis it at Burstall because it is the only site that switched from unflooding to flooding, and the other threes are consistently either non-flooding or consistently flooding. This analysis allows us to directly assess the influence of flooding on model performance.

Lastly, we expanded the descriptions of the study sites by including additional information on site characteristics, meteorological conditions, water table dynamics, and the procedures used to reconstruct the daily measurement-based T:ET values. We also substantially expanded the discussion to examine the factors controlling ET partitioning performance, including site characteristics, canopy structure, ecosystem heterogeneity, eddy covariance measurement configuration, and the assumptions underlying the different partitioning approaches. In particular, we added a new discussion section specifically addressing the potential effects of moss cover and flooding on the performance of the evaluated ET partitioning methods.

These revisions resulted in several new findings while reinforcing our original conclusions.

- (1) None of the evaluated ET partitioning approaches consistently outperformed the others across study sites or under both flooded and non-flooded conditions.

- (2) Regardless of whether the independent measurement-based T:ET or the all-method ensemble was used as the benchmark, no individual method consistently performed better than simply using the observed mean as the predictor.
- (3) None of the three methodological groups (high-frequency EC-based, ecosystem carbon-water coupling, or machine learning approaches) demonstrated a clear overall advantage over the others.
- (4) The evaluation results are influenced by the choice of reference dataset, which highlights the importance of benchmark selection when assessing ET partitioning performance.
- (5) The performance of the methods is influenced by the combined effects of multiple environmental and methodological factors, including moss cover, flooding conditions, canopy structure, ecosystem heterogeneity, eddy covariance measurement configuration, data quality, and the assumptions underlying each partitioning approach, rather than by moss cover or flooding alone.

Overall, we believe that these substantial revisions considerably strengthen the manuscript by providing a more comprehensive and rigorous evaluation of ET partitioning approaches in our study sites. Below are our point-by-point responses to the reviewers' comments. We have updated our previous responses to reflect the revisions made in the manuscript. Unless otherwise noted, all line numbers throughout our responses refer to the tracked-changes version of the revised manuscript.

Once again, we sincerely thank the Editor and the Reviewers for their thoughtful comments, which have greatly improved the quality of this manuscript. We hope that the revised manuscript satisfactorily addresses all concerns and will be suitable for publication.

Reviewer #1:

Review of "How well do evapotranspiration partitioning approaches perform in moss-covered wetlands?"

This paper delves into an important topic of evapotranspiration (ET) partitioning in moss-covered wetlands, comparing different methods against "direct" measurements. The topic is relevant and the manuscript is generally well written and easy to follow. To my knowledge, this is the first manuscript that compares such a wide range of ET partitioning methods, including high-frequency EC and GPP-informed methods. While the study sites are not ideal given the high heterogeneous conditions, I believe it can still provide useful insights on the performance of the different methods. My only concern about the study sites is the presence of flooded areas, which might complicate the interpretation of the results or even invalidate some of the methods. Overall, I think the manuscript has merit and can be published after addressing the comments below.

Line 106: Given that was developed specifically for flooded ecosystems, it is almost required to be included in this analyses. I would strongly encourage the authors to contact the group by Eichelmann et al. (2022) to get help using their code.

Response: Thank you very much for your positive comments and constructive suggestions. Following Dr. Eichelmann's recommendation, we incorporated the machine learning framework developed by Stapleton et al. (2022), which is based on the same ET partitioning assumptions but employs different machine learning algorithms and a machine learning-based feature selection procedure. We evaluated the best-performing machine learning model identified for each study site, and the results are presented in Figure 4 and Table 5. Based on both the coefficient of determination (R^2) and the Nash-Sutcliffe model efficiency coefficient (NSE), we found that the machine learning approach does not consistently outperform the other ET partitioning approaches across the four study sites or under both flooded and non-flooded conditions. Moreover, in most cases, its predictions perform no better than simply using the observed mean as the predictor. The potential reasons for these findings are discussed in Section 3.3.

Reference:

Stapleton, A., Eichelmann, E., Roantree, M., 2022. A framework for constructing machine learning models with feature set optimisation for evapotranspiration partitioning. *Applied Computing and Geoscience*, 100105. <https://doi.org/10.1016/j.acags.2022.100105>.

Line 128: study sites

One important aspect of the high-frequency partitioning methods is that the ground is a source of CO₂ (respiration) and source of water vapor (evaporation). Moss photosynthesis at the ground level might not necessarily invalidate the methods if we consider the net effect ($\text{surface_level_respiration} - \text{moss_P}$), as long as the CO₂ signal from the surface is positive and well mixed before it reaches the canopy top and sensor (i.e., different from being fully mixed with the plant canopy signal). What is unclear to me is the role of the water in flooded parts. If water is a sink of CO₂, then the net CO₂ flux from surface level is negative, and none of these methods would make sense given the decoupling between water vapor and CO₂.

Response: We thank the reviewer for raising this important point regarding the effects of flooding on high-frequency eddy covariance (EC)-based ET partitioning methods. In the revised manuscript, we have explicitly examined the influence of flooding on the performance of the evaluated methods.

We agree that flooding can weaken or obscure surface CO₂ signals and reduce the quality of EC measurements. However, we do not believe that flooding renders these methods, particularly the high-frequency conditional sampling approaches, entirely inapplicable. For example, the high-frequency methods still produced reasonable T:ET predictions at the continuously flooded Poplar site (Fig. 4). Nevertheless, flooding did deteriorate the performance of some methods. To quantify this effect, we recalculated the R² and NSE values for Burstall after excluding the flooded period. We found that CEA, MREA, and PP showed slight improvements in R².

In terms of the effect of moss cover, we would like to clarify that the assumption of a positive surface CO₂ signal may not always hold in moss-covered wetlands. Strong moss photosynthesis can result in net CO₂ uptake at the surface, which means that surface-level respiration minus moss photosynthesis can be negative. Evidence for this behavior is shown in Walker et al. (2017), where chamber measurements on Sphagnum–peat columns (their Fig. 4) reveal clear transitions between positive and negative hourly CO₂ fluxes, as well as sustained daily net CO₂ uptake from mid-June to mid-July. These observations indicate that the presence of a moss layer can challenge the applicability of high-frequency EC based ET partitioning methods that assume a net positive surface CO₂ signal.

For our study sites, direct CO₂ measurements during the study periods were unfortunately not available. However, existing evidence suggests that negative surface CO₂ flux signals are plausible, particularly at the Poplar site. Following a wildfire in 2016, a chamber-based ground CO₂ flux study conducted in 2017 at the Poplar site measured moss-dominated plots and showed that the unburned ground surface acted as a net CO₂ sink until early August (van Beest, 2019, Fig. 3-1). This further supports the possibility that surface CO₂ uptake may occur at some of our sites.

In response to your comment, we have substantially expanded the discussion of both flooding and moss cover throughout the revised manuscript. These topics are addressed primarily in Sections 3.3 and 3.5,

References:

Walker, A. P., K. R. Carter, L. Gu, P. J. Hanson, A. Malhotra, R. J. Norby, S. D. Sebestyen, S. D. Wullschleger, and D. J. Weston (2017), Biophysical drivers of seasonal variability in *Sphagnum* gross primary production in a northern temperate bog, *J. Geophys. Res. Biogeosci.*, 122, 1078–1097, <https://doi.org/10.1002/2016JG003711>.

Van Beest, C. Deeper Burning Increases Available Phosphorus, Promotes Moss Growth, and Carbon Dioxide Uptake in a Fen Peatland One-Year Post-Wildfire in Fort McMurray, AB. Master's thesis, University of Waterloo, 2019.

Line 225: Results and discussion

I agree with the authors when they say that direct measurements cannot be directly compared to any of the approaches since they all use EC fluxes, covering a much larger and heterogeneous footprint. I would also add that even under ideal homogeneous conditions, direct measurements (sap flow, chambers, etc) still suffer from other limitations and cannot be considered as the "truth", but as proxies.

Response: Thank you for this important clarification, which we fully agree with. In the manuscript, we already note several limitations of direct measurements, including flux chambers, porometers, and lysimeters (Lines 269 to 297).

To provide a more comprehensive evaluation, we therefore used two complementary reference datasets: (1) the independent measurement-based T:ET reconstructed from field measurements and (2) the ensemble mean of the ET partitioning approaches (Lines 355 to 369). Interestingly, we found no significant correlation between these two reference datasets, and the mean T:ET of the all-method ensemble was 10.4% higher than that of the measurement-based estimates. These findings highlight that the choice of reference dataset can substantially influence the evaluation results and further emphasize the uncertainty associated with assessing ET partitioning performance in wetlands.

Fig 1 and 2: Are "measurement-based" data an average including all "direct measurements"? Is there any measure of variance, say, across chambers/plants (for sap flow) or maybe across methods?

Response: Thank you for these thoughtful questions. To answer the first question briefly, the measurement-based data are not simple averages of all direct measurements.

For the Sibbald, Burstall, and Bonsai sites, the measurement-based T:ET time series are derived from a site-calibrated and validated Shuttleworth–Wallace (S–W) model rather than direct averaging of ground evaporation measurements. This approach was necessary because micro-lysimeters provide averaged daily evaporation over multi-day periods and because eddy-covariance ET measurements at these sites include intermittent data gaps. To address this, we developed a surface resistance formulation that explicitly accounts for moss and litter effects on ground evaporation (Wang et al., 2023) and incorporated it into the S–W model.

The S–W model performance was evaluated against eddy-covariance measurements, with linear regressions of modelled versus observed latent heat flux showing slopes ranging from 0.87 to 1.16, intercepts between -18.5 and 3.8 W m^{-2} , and R^2 values greater than 0.78 across the three sites. Modelled ground evaporation was further compared with microlysimeter-based measurements conducted over peat-only surfaces, peat surfaces covered with moss, and peat surfaces covered with both moss and litter, under both open conditions and a willow canopy. Modelled evaporation was generally slightly lower than the measurements, which is expected given that lysimeter experiments were conducted under controlled conditions with sedge removal and therefore reduced shading. Importantly, modelled evaporation fell within the observed

measurement range and reproduced similar temporal variability. For conditions consistent with the S–W model framework, namely litter- and moss-covered surfaces under canopy, the agreement between modelled evaporation and measurements was particularly strong, with a linear regression given by $\lambda E_{\text{modelled}} = 0.91 \cdot \lambda E_{\text{measured}} - 4.46 \text{ (W m}^{-2}\text{)}$ and $R^2 = 0.85$. These results support the use of the S–W model to construct T:ET time series at the three sites. Additional validation details are provided in Wang (2025, pp. 110–112), and key procedures and results will be summarized in the revised manuscript.

At the Poplar site, the measurement-based T:ET time series was constructed differently from the other three sites, as this site includes direct eddy-covariance ET measurements and upscaled sap flow measurements for dominant tree and understory species, informed by forest inventory data and LiDAR-based vegetation classification, as well as ground-based chamber flux measurements (Gabielli, 2016). Ground evaporation was calculated as the residual (ET minus upscaled site transpiration (T)) and evaluated against chamber measurements dominated by moss. This comparison yielded a regression slope of 0.96, an intercept of 0.08 mm h^{-1} , and an R^2 of 0.26. Because both T and ET were measured at the same timestep, whereas chamber measurements are instantaneous and inherently variable, we used the sap-flow-derived transpiration together with eddy-covariance ET to construct the Poplar T:ET time series. We acknowledge the relatively low R^2 , which reflects the scatter in the chamber measurements, but the near-unity slope and small intercept indicate good overall agreement. On this basis, we consider the resulting T:ET time series to be a reasonable representation for this site.

Overall, constructing measurement-based T:ET estimates is non-trivial and that these data should not be considered “truth,” but rather proxies with inherent uncertainty. We therefore appreciate the reviewer’s comment regarding the interpretation of “truth” data.

In response, we have provided additional technical details on the field measurements (Section S3 of the supplementary Material) and the procedures used to reconstruct the daily measurement-based T:ET time series (section 2.4), and the validation of the measurement-based T:ET (Section 2.4 and Section S3 of the supplementary Material).

Regarding the second question, the table below summarizes errors and uncertainties associated with flux measurements relevant to this study. This table is now included as Table S4 in the revised manuscript.

Table 1. Examples of Errors and Uncertainties of Flux Measurements

Flux contributions	Method	Source of errors	Normalized Root Mean Square Error (%)	Reference
ET	EC	One-point Uncertainty	8.3	Kroon et al. 210
T	Sap flow system (stem heat balance method)	System measurement resolution, Steady-state assumption, Heat-storage error	15.0	Groot and King, 1992 Shackel et al., 1992 Grime et al., 1995 Perämäki et al., 2001
ET (ground)	Chamber	Lower Q*, Altered Microclimate	6.6	Reicosky et al., 1983 McLeod et al., 2004 Hamel et al., 2015
E	Micro-lysimeter	Soil disturbances, Handling and weighing	17.8	Calculated using micro-lysimeter measurements from this study

(Note: Table adapted from Gabrielli 2016).

References:

Gabrielli, E. C.: Partitioning Evapotranspiration in Forested Peatlands within the Western Boreal Plain, Fort McMurray, Alberta, Canada, Master's thesis, Wilfrid Laurier University, 2016.

Grime, V.L., Morrison, J.I.L., Simmonds, L.P. 1995. Including heat storage term in sap flow measurements with the heat storage method. *Agricultural and Forest Meteorology*, 74: 1-25.

[https://doi.org/10.1016/0168-1923\(94\)02187-O](https://doi.org/10.1016/0168-1923(94)02187-O).

Groot, A., King, K.M. 1992. Measurement of sap flow by heat balance method: numerical analysis and application to coniferous seedlings. *Agricultural and Forest Meteorology*, 59: 289-308. [https://doi.org/10.1016/0168-1923\(92\)90098-O](https://doi.org/10.1016/0168-1923(92)90098-O).

Hamel, P., Mchugh, I., Coutts, A., Daly, E., Beringer, J., Fletcher, T.D. 2015. Automated Chamber System to Measure Field Evapotranspiration Rates. *Journal of Hydrologic Engineering*, 20(2): 04014037. [https://doi.org/10.1061/\(ASCE\)HE.1943-5584.0001006](https://doi.org/10.1061/(ASCE)HE.1943-5584.0001006).

Kroon, P.S., Hensen, A., Jonker, H.J.J., Ouwersloot, H.G., Vermeulen, A.T., Bosveld, F.C. 2010. Uncertainties in eddy covariance flux measurements assessed from CH₄ and N₂O observations. *Agricultural and Forest Meteorology*, 150: 806-816.

<https://doi.org/10.1016/j.agrformet.2009.08.008>.

McLeod, M.K., Daniel, H., Faulkner, R. Murison, R. 2004. Evaluation of an enclosed portable chamber to measure crop and pasture actual evapotranspiration at small scale. *Agricultural Water Management*, 67: 15-34. <https://doi.org/10.1016/j.agwat.2003.12.006>.

Perämäki, M., Vesala, T., Nikinmaa, E. 2001. Analyzing the applicability of the heat balance method for estimating sap flow in boreal forest conditions. *Boreal Environment Research*, 6: 29-43. <https://doi.org/10.60910/4tn6-73en>.

Reicosky, D.C., Sharratt, B.S., Ljungkull, J.E., Baker, D.G. 1982. Comparison of Alfalfa Evapotranspiration Measured By A Weighing Lysimeter And A Portable Chamber. *Agricultural Meteorology*, 28: 205-211. [https://doi.org/10.1016/0002-1571\(83\)90026-2](https://doi.org/10.1016/0002-1571(83)90026-2).

Wang, Y., Petrone, R. M., and Van Huizen, B.: The dependence of evaporative efficiency of vegetated surfaces on ground cover mass fractions in vegetated soils in mesic ecosystems, *Hydrological Processes*, 37, e15036, <https://doi.org/10.1002/hyp.15036>, 2023.

Wang, Y.: Uncovering the understudied role of microtopography and ground cover on evapotranspiration partitioning in high-elevation wetlands in the Canadian Rocky Mountains, Doctoral dissertation, University of Waterloo, 2025.

Line 272: Is it possible that data quality was worse when $T/ET < 0.5$?

Response: Thank you for this insightful question. Yes, it is possible that data quality was poorer during periods when $T/ET < 0.5$. We found that these low T/ET values generally coincide with days characterized by low net radiation, low air temperature, and high soil moisture. This does not necessarily imply flooded conditions, as the Poplar site was consistently flooded throughout the study period while maintaining T/ET values above 0.5.

Under such conditions, several data quality issues may arise. For eddy-covariance measurements, high moisture availability combined with weak vertical mixing can lead to surface evaporation that is not efficiently transported upward. As a result, latent heat flux measurements can become

noisy and are often underestimated. This suppressed vertical mixing may also introduce bias in high-frequency EC-based ET partitioning methods.

Similarly, low net radiation and temperature generally reduce transpiration rates, leading to low sap flow magnitudes. When sap flow signals are weak, instrument noise and relative measurement uncertainty become more influential, potentially affecting transpiration estimates.

These data quality considerations are discussed throughout the Results and Discussion sections of the revised manuscript.

Line 300: Conclusion

I would refrain from talking about more or less "accurate" results. One of the main challenges of flux partitioning is that we really do not know the true values of any of the components, and while direct measurements are good proxies of trends, they have limitations and might even required some level of parameterization such as sap flow measurements.

Response: Thank you for this helpful suggestion. We agree that referring to results as more or less “accurate” is not appropriate given the inherent uncertainties in flux partitioning and the proxy nature of both model- and measurement-based estimates. we have extensively revised the Conclusions to avoid references to "accuracy" and instead discuss the results in terms of relative performance, consistency, and agreement among the evaluated approaches.

Reviewer #2:

Wang et al., Technical note: How well do evapotranspiration partitioning approaches perform in moss-covered wetlands?

In the interest of transparency, I would like to disclose that some of the work discussed in this manuscript and my review relates to some of my own publications (e.g. Eichelmann et al., 2020; Stapleton et al., 2022).

General remarks:

In this manuscript, Wang et al. conduct a comparative study to evaluate nine different evapotranspiration partitioning tools/models for eddy covariance data at four moss-covered wetland sites over one growing season at each site (total of four site-years).

They conclude that none of the tools captures all the temporal and spatial variability, but that in combination they can provide useful information on T:ET estimates for these sites.

Overall the manuscript investigates an interesting and relevant subject. Cross-site and cross-tool comparisons of eddy covariance based ET partitioning methods are still very rare and are much needed. This is especially true for some of the more unique and challenging ecosystems such as those investigated in this manuscript, where models aren't routinely evaluated. The manuscript is well written and the science is generally sound.

That all being said, I think that the manuscript at the moment, unfortunately, falls short of being actually useful for other researchers in a concrete way. The overall conclusions are quite broad and vague and don't go much beyond what existing knowledge already told us.

Response: Thank you for your thoughtful assessment of our work. We appreciate your recognition of the relevance of cross-site and cross-method evaluations of ET partitioning approaches, particularly for moss-covered wetlands where such comparisons are still rare.

We have revised the Results and Discussion to more explicitly identify the conditions under which specific methods perform well or poorly, highlight common failure modes, and clarify the strengths and limitations of each approach in wetland settings. We also refined the Conclusions to move beyond general statements and provide more targeted recommendations for researchers selecting ET partitioning methods in moss-covered and hydrologically complex ecosystems.

We believe these revisions address your concern and substantially improve the practical usefulness of the manuscript. Detailed responses to your specific comments and suggestions are provided below.

I think there are two main reasons for this:

- 1) My main concern is the lack of quantitative assessment/comparison of the ET partitioning performance. I acknowledge that it is extremely difficult if not impossible to get 'ground truth' data to validate model performance against, and commend the authors on their

efforts to obtain some validation data for their sites. While I agree with the authors that the micro-lysimeter and Shuttleworth-Wallace model validation data is not a true ground truth and any comparison between them has to be taken with a grain of salt, I still think that there should be a quantitative evaluation of the model performance, rather than just a qualitative assessment. At the end of the day, there is a set of independent data, which can be used as a baseline for comparison across the models. While we might not expect a perfect fit with these baseline data, we can still get quantifiable differences across models for a more robust discussion. See my specific comments further below for more detail on this.

Response: Thank you for this constructive and important suggestion. We agree that a more quantitative evaluation would strengthen the manuscript and improve its usefulness for other researchers.

In response to this comment, we substantially expanded the performance evaluation in the revised manuscript. Specifically, we introduced the Nash-Sutcliffe model efficiency coefficient (NSE) in addition to the coefficient of determination (R^2), evaluated the methods against both the independent measurement-based T:ET and the all-method ensemble mean, examined the agreement among methods using Spearman correlation analysis, and investigated the influence of flooding by comparing model performance before and after the flooded period at Burstall. We also expanded the discussion to interpret these new analyses and to identify the environmental and methodological factors influencing ET partitioning performance.

- 2) I also would have liked to see a bit more detail on the environmental/meteorological conditions at these sites across the study periods and how they could be influencing some of the observed patterns (or lack thereof). There are no high temporal resolution data presented (only an overall range or average for each site across the measurement period) on water table depth, temperature (soil/water/air), VPD, or precipitation. As the authors state, some of these parameters exert strong influences on biosphere-atmosphere exchange of water, and some of the tools/models are better or worse at incorporating these impacts. Providing some of these meteorological data alongside the partitioning results and/or evaluating under what conditions certain tools perform better/worse would

improve the manuscript and make it more useful for other researchers to help decide which tools might be best in their context. But I recognize that the limited data availability might restrict how much can be learned from this.

Response: Thank you for this constructive suggestion. We agree that providing more detail on environmental and meteorological conditions, and explicitly linking them to model performance, would improve the manuscript's usefulness.

In response, we have added a new figure presenting the temporal dynamics of key meteorological variables and water table depth for all four study sites (Section S2 of the Supplementary Material, Fig. S4), together with additional descriptions of the site characteristics and environmental conditions. We have also substantially expanded the discussion to examine how factors such as canopy structure, canopy density and openness, ecosystem heterogeneity, meteorological conditions, data quality, flooding, and moss cover may influence the performance of the evaluated ET partitioning approaches.

We also attempted to investigate the influence of flooding by comparing model performance before and after the flooding period at Burstall, which was the only site that transitioned from non-flooded to flooded conditions during the study period. However, because the four study sites differ substantially in vegetation, hydrology, canopy structure, and environmental conditions, and because flooding occurred at only one site, we were unable to conduct a rigorous statistical analysis to isolate the effects of individual environmental variables on model performance. We have therefore interpreted these factors qualitatively rather than quantitatively. We hope that the additional analyses, site information, and expanded discussion adequately address the reviewer's concerns.

Overall this study is a very worthwhile effort, but requires a bit more quantitative analysis to realize its full potential.

Response: Thank you for this encouraging assessment. Thank you for this encouraging assessment. We appreciate your positive comments and agree that a more quantitative evaluation strengthens the manuscript. In response, we have substantially expanded the quantitative analyses, as summarized in our overall response above. We believe these additions have

considerably strengthened the manuscript and improved its usefulness for researchers interested in ET partitioning in wetland ecosystems.

Specific comments:

Introduction: Well written, concise introduction with a clear justification for the study and sufficient background information provided. Relevant literature is cited, but it might be interesting to have a look at Speranskaya et al. (2024) also.

Response: Thank you for the positive feedback and for the suggested reference. We have incorporated Speranskaya et al. (2024) into the Introduction.

Methods

Line 106:

While the original ANN based partitioning code from our 2020 paper unfortunately cannot be made publicly available, a subsequent publication using different machine learning algorithms, but based on the same underlying concept, has since been published (Stapleton et al., 2022). The python code for this is publicly accessible and linked in that publication. It might be an interesting comparison to include, given that it works on evaporation prediction and doesn't require ecosystem carbon-water coupling. As far as I can see all other non-high-frequency based methods in this manuscript rely on carbon-water coupling, so it could be interesting to compare with the machine learning based method and see if different patterns emerge.

Response: Thank you for this constructive and helpful suggestion. Reviewer #1 also recommended that including the machine learning framework developed by your group would substantially strengthen the manuscript. We have therefore incorporated this framework into our evaluation. Surprisingly, however, the best-performing machine learning models did not consistently outperform the other ET partitioning approaches across the four study sites or under flooded conditions. The potential reasons for this are discussed in Section 3.3.

Line 149, Table 2: Please add information on the measurement period for each site (start and end date)

Response: Thank you for this helpful suggestion. We have added the start and end dates of the measurement period for each study site to Table 2 in the revised manuscript. In addition, we have included the measurement periods for the ground evaporation in Table S3.

Line 161: Maybe add the Pastorello et al. (2020) reference for the Fluxnet data processing protocol

Response: Thank you for pointing out this oversight. We have added Pastorello et al. (2020) to the revised manuscript.

Line 173-175: I understand that the detailed upscaling procedures have been described elsewhere, but I still think the authors should provide a short (couple of sentences) summary of the main steps to help the reader get a sense of what is involved here without having to dig up the other publications and reading up on the detail.

Response: Thank you for this helpful suggestion. Reviewer #1 raised similar questions regarding the construction of the measurement-based T:ET time series, and we agree that providing a brief summary of the upscaling procedure would improve clarity. In response, we have added a detailed description of the reconstruction procedure in Section 2.4 and provided additional technical details in Section S3 of the Supplementary Material.

Results and Discussion

Figure 1 and 2: For Bonsai, it looks as though there are more ‘Measurement-based’ black circles in Figure 2 compared to Figure 1, especially in the late growing season after Aug 15, but there are also some of the higher values missing earlier in the growing season (e.g. a measurement based value close to 1 around August 1). Why are some of those measurements omitted from Figure 1?

Response: Thank you for pointing out this inconsistency. You are correct that the measurement-based data points should be identical between Figures 1 and 2 for the same study site. Upon carefully reviewing our data processing workflow, we identified two issues that contributed to this discrepancy.

First, some ET partitioning methods only produce predictions for dates with valid model outputs. During data processing, we used left joins in R to merge datasets by date. A left join retains all dates from the first dataset and only adds matching observations from the second dataset. Consequently, any dates that are absent from the first dataset are excluded from the merged dataset, even if they are present in the second dataset. Because the Group 1 and Group 2 figures were generated using different datasets as the starting point for the merge, a small number of dates were inadvertently excluded in one figure but retained in the other. As a result, the measurement-based T:ET time series was not represented consistently between the two figures. We have corrected the data merging procedure by ensuring that all datasets are merged using a common reference date sequence, so that the measurement-based time series is represented consistently across all figures. We found that this issue affected not only Bonsai but also a small number of data points at the other study sites.

Second, we identified an issue in our precipitation filtering procedure. In the original analysis, daytime observations were selected before applying the precipitation filter. Consequently, some days with precipitation occurring only during the nighttime were mistakenly retained, even though they should have been excluded. We have corrected the order of the filtering steps so that precipitation filtering is applied before selecting daytime observations. This issue has now been resolved throughout the dataset.

We thank the reviewer for bringing these issues to our attention. Correcting them has improved the consistency of the figures and the robustness of the revised analyses.

There doesn't seem to be a lot of temporal variability in the measured T:ET values and most of the models don't capture the temporal variability well. I am wondering, are any of these models actually better than just assuming a constant T:ET ratio? I think it would be good to add that in for comparison.

Response: Based on the NSE results, we found that none of the evaluated ET partitioning approaches consistently outperformed simply using the mean of the reference dataset as the predictor, regardless of whether the independent measurement-based T:ET or the all-method ensemble was used as the reference dataset. This finding has been added to the Results and Discussion sections of the revised manuscript and further emphasizes the challenges of ET partitioning in moss-covered wetlands.

I would also like to see a bit more of a quantitative analysis on the performance of the models. At the moment, everything is based on reading off approximate performance from the graphs and is quite qualitative, e.g. lines 290-292: “Both CP and CWSC [...] produced a narrower spread of T:ET estimates than uWUE ...” I understand that the ‘ground truth’ data based on Micro-lysimeters and the Shuttleworth-Wallace model isn’t really a ground truth and fully comparable 1:1, but I still think some level of quantitative analysis would be beneficial. While we maybe wouldn’t expect the graphs in Fig 3 to fall on the 1:1 line, even just providing R² values (within and across sites) for how well the models track temporal and spatial patterns would be interesting and will make the resulting discussion more robust. The authors may also want to explore other measures of goodness of fit, which might be more appropriate in this particular situation (e.g. Nash–Sutcliffe coefficient).

Response: Thank you for this constructive suggestion. We agree that a more quantitative evaluation strengthens the manuscript. In response, we incorporated both the coefficient of determination (R²) and the Nash-Sutcliffe model efficiency coefficient (NSE), together with several additional quantitative analyses. To avoid repetition, these major additions are summarized in our overall response above and are described in detail in the revised manuscript.

Line 296-298: I think it is a bit of a stretch to say these models still offer valuable insights. Just from looking at the graphs in Fig 3, it seems to me that assuming a constant T:ET ratio based on an approximate global average of 0.6 (e.g. Wei et al., 2017; Liu et al., 2022) would perform just as well across sites and in some cases also within sites (especially at Brustall). So this statement really needs to be backed up with some quantitative analysis.

Response: Thank you for this helpful comment. We agree that our original conclusion was not sufficiently supported by quantitative evidence. In the revised manuscript, we have substantially revised the Results, Discussion, and Conclusions to reflect the new quantitative analyses. Based on the NSE results, we found that none of the evaluated ET partitioning approaches consistently outperformed simply using the mean of the reference dataset as the predictor. Accordingly, we have removed statements implying that the evaluated methods provide more valuable insights than this simple benchmark and have reframed our conclusions in terms of the relative performance, consistency, and limitations of the different approaches.

Summary and closing thoughts

Line 305: Again, here it says that the models captured similar temporal patterns of T:ET. I'm not sure what this statement is really based on given that most sites don't exhibit much temporal variation at all, and and at the one site that does (Brustall) only one out of 9 models manages to capture that temporal trend.

Response: Thank you for pointing this out. We agree that this statement was not sufficiently supported by the results. In the revised manuscript, we have removed this generalization and revised the discussion based on the quantitative analyses. As shown by the R^2 results (Table 2 to 5), only a few methods were able to reproduce the temporal variability of the reference T:ET, while most approaches showed limited ability to capture temporal variation. We have revised the text accordingly to more accurately reflect the results.

References:

Liu, Y., Zhang, Y., Shan, N., Zhang, Z., & Wei, Z., 2022. Global assessment of partitioning transpiration from evapotranspiration based on satellite solar-induced chlorophyll fluorescence data. *Journal of Hydrology*, 612. <https://doi.org/10.1016/j.jhydrol.2022.128044>

Pastorello, G., Trotta, C., Canfora, E. et al., 2020. The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. *Sci Data* 7, 225. <https://doi.org/10.1038/s41597-020-0534-3>

Speranskaya, L., Campbell, D. I., Lafleur, P. M., and Humphreys, E. R., 2024. Peatland evaporation across hemispheres: contrasting controls and sensitivity to climate warming driven by plant functional types. *Biogeosciences*, 21, 1173–1190. <https://doi.org/10.5194/bg-21-1173-2024>

Stapleton, A., Eichelmann, E., Roantree, M., 2022. A framework for constructing machine learning models with feature set optimisation for evapotranspiration partitioning. *Applied Computing and Geoscience*, 100105. <https://doi.org/10.1016/j.acags.2022.100105>

Wei, Z., Yoshimura, K., Wang, L., Miralles, D. G., Jasechko, S., & Lee, X., 2017. Revisiting the contribution of transpiration to global terrestrial evapotranspiration. *Geophysical Research Letters*, 44(6), 2792–2801. <https://doi.org/10.1002/2016GL072235>

Public justification (visible to the public if the article is accepted and published):

Thank you for your comprehensive and thorough revision of your manuscript. The main concerns of the reviewers have been addressed. Additional partitioning methods based on machine learning and quantitative comparison of methods have been added.

In my opinion, the following points have to be improved first (minor revisions) before this manuscript will be acceptable for publication.

Response: Thank you very much for recognizing that we have addressed the major concerns raised by the reviewers. We also sincerely appreciate your additional suggestions for further improving the readability and presentation of the manuscript.

Following your comments, we have carefully revised the manuscript throughout to improve its clarity, readability, and overall presentation. Below, we provide a detailed, point-by-point response to each of your comments and describe the corresponding revisions made in the manuscript.

We hope that these changes have further improved the quality and readability of the manuscript, and we sincerely thank you for your time and constructive feedback.

- Please double check for the entire manuscript, where you can use defined abbreviations, such as ET, T and E, and where you should write out the words (e.g., in lines 73-78 you use ‘transpiration’ and ‘evaporation’, where you could also just use ‘T’ and ‘E’). For instance, use no abbreviations in titles or at the beginning of sentences. Check if you have defined all abbreviations in the text, when you use them the first time.

Response: Thank you for your helpful suggestion. We have carefully reviewed the use of abbreviations throughout the entire manuscript and Supplementary Material, including the abstract, main text, section headings, figures, tables, and their titles and captions. We have revised the manuscript throughout to ensure that all abbreviations are defined at their first use and used appropriately thereafter. We have also ensured that abbreviations are not used in titles or at the beginning of sentences.

- line 68: please rephrase, maybe as follows: “In fact, it cannot be guaranteed that these methods are suitable for wetlands, ...”

Response: Revised as suggested. (Line 71 in the tracked-changes version).

- line 85: It should say “...relationships between carbon uptake and water...”

Response: revised as suggested. (Line 90 in the tracked-changes version).

- lines 95-96: It should say: “...reported moss contributions to cumulative community-level photosynthesis of about 1.5%, 35.2% and 41.2% in ...”

Response: Revised as suggested. (Line 100 in the tracked-changes version).

- Table 1: The definition of ‘Ph’ is missing (used in the description of PP method). Please add in Notes under table or don’t use it.

Response: Thank you for pointing this out. Based on the PP method (Pérez-Priego et al. (2018)), the “Ph” is the EC-inferred photosynthesis rate expressed as the molar flux density ($\mu\text{mol CO}_2 \text{ m}^{-2} \text{ s}^{-1}$). We revised it to “EC-inferred photosynthesis rate” for better clarity. (Table 1 in the tracked-changes version).

- line 200: ‘Poplar’ with capital ‘P’. Please check entire manuscript for correct study site names.

Response: Thank you for your suggestion. We have revised this as suggested and carefully checked the entire manuscript and supplementary material to ensure that all study site names are used consistently and correctly throughout.

- Table 2: Please change the description of ‘vascular vegetation’ from complete sentences to bullet points. Maybe, change the format of the table to improve readability. Use one column for each sites and all the characteristics as rows (thus, transpose the table).

Response: Thank you for your suggestion. We have reformatted the description of vascular vegetation using bullet points and transposed Table 2 so that the study sites are presented as columns and the site characteristics as rows.

- line 251: Check the usage of dashes and hyphens in entire manuscript (e.g. for ‘S-W model’)

Response: Thank you for your suggestion. We have carefully reviewed the use of dashes and hyphens throughout the manuscript and supplementary material and revised them to ensure consistency with the journal's style. For example, "S-W model" has been corrected to "S–W model", and "Priestley-Taylor equation" has been corrected to "Priestley–Taylor equation."

- lines 261-268: Can you add for this paragraph a reference or a figure to the Supplementary material? Or just state “(not shown)” somewhere.

Response: Thank you for your helpful suggestion. We have added the comparison and fitted linear regression as Fig. S6 in the Supplementary Material and referred to this figure in the revised manuscript. (Line 490 in the tracked-changes version).

- lines 326-332: Could you please add your used notations for the different ensemble means in this paragraph: all-method ensemble, HF ensemble, CW ensemble, and ML ensemble?

Response: Thank you for your helpful suggestion. We have added the notations for the different ensemble means in the revised manuscript as follows: “For each day, the overall ensemble mean was calculated as the average of all available model predictions of T:ET and is hereafter referred to as the all-method ensemble. In addition, ensemble means were calculated separately for each method group, referred to as the HF ensemble (Group 1 ensemble), CW ensemble (Group 2 ensemble), and ML ensemble (Group 3 ensemble). These three group-specific ensembles were evaluated against both the measurement-based T:ET time series and the all-method ensemble to assess whether any group of ET partitioning approaches consistently outperformed the others.” (Lines 383 to 388 in the tracked-changed version).

- Fig. 1, 2, and 4: May it be possible to add some kind of mark (e.g., a horizontal bar) in the panels to indicate flooded periods at Poplar and Burstall?

Response: Thank you for your helpful suggestion. We have added cyan bars at the bottom of the Burstall and Poplar panels in Figs. 1, 2, and 4 to indicate the flooded periods. The corresponding figure captions have also been revised to describe these additions.

- line 380: It should say: “Note: R² values ...”. Adjust for all related tables.

Response: Thank you for your suggestion. To improve clarity, we have revised the note in all related tables from "Values greater than 0.5 are shown in bold" to "Any R² or NSE value greater than 0.5 is shown in bold."

- line 386: I guess it should say “Klosterhalfen et al. (2019b)...”

Response: Thank you for pointing this out. The citation has been corrected to "Klosterhalfen et al. (2019b)". (Line 458 in the tracked-changed version).

- Could you please improve the resolution of all figures in the supplementary material.

Response: Thank you for your helpful suggestion. We have improved the resolution of all figures in the Supplementary Material to 600 dpi. We also increased the font sizes of the axis titles, tick labels, panel titles, and legends to improve readability. The only exception is Fig. S3 (site photographs), which was already prepared using the highest-resolution images available, so no changes were made to that figure.

- Table S4: The given RMSE is calculated between what (e.g., measurements vs model)? Could you add a brief info in the table caption?

Response: Thank you for pointing this out. The RMSE reported in Table S4 is the mean test-set root mean squared error of the selected best model across the cross-validation folds. Specifically, using 10-fold cross-validation, the nighttime data were randomly divided into 10 subsets. In each iteration, the model was trained on 9 folds and evaluated on the remaining held-out fold. The

root mean squared error was calculated for each held-out fold, and the reported RMSE is the mean across all 10 folds. Following Stapleton et al. (2022), RMSE was used for feature selection, but not for selecting the best model. The best model was selected based on the highest adjusted R^2 , as formulated in Stapleton et al. (2022). We have added this information to the Table S4 caption for clarification.