

Response to comments on Manuscript egusphere-2026-2687 (RC1)

Title: Physically constrained multi-task learning for hourly joint estimation of evapotranspiration and transpiration from sparse sap-flow data

Authors: Hongyang Fu, Yuanyuan Yang*, Dengfeng Liu*, Qiang Li, Zhihua He, Huimin Lei, Mohd Yawar Ali Khan, and Fiaz Hussain

Manuscript ID: egusphere-2026-2687

Citation: <https://doi.org/10.5194/egusphere-2026-2687-RC1>

Reply on RC1:

Many thanks for taking the time and effort to review our paper. All comments from Referee #1 are addressed below with point-by-point responses.

For better readability, our replies are prefixed with “R/”. The original reviewer comments are prefixed with “RC-n/” and are presented in bold. The revisions to be added into the revised manuscript are highlighted in red. The quoted content is displayed in italics. All references cited in this response are listed at the end of the reply. Line numbers from the unrevised manuscript are provided where relevant.

Point-to-point response:

RC1-0/ The paper presents and applies a methodology for the joint estimation of hourly ET and T (transpiration) by training various machine learning models (based on neural networks and decision trees) on measurements of ET (eddy covariance) and T (sap flow data). Joint learning is enabled by using a multi-objective calibration function, using model architectures that share components for predicting ET (task 1) and the ratio T/ET (task 2), combined with enforcing the constraint $T < ET$. This setup allows for data on T to inform ET estimates, and vice versa. The paper investigates to what extent this joint estimation improves temporal prediction of ET and T at individual sites ('temporal generalization') and spatial prediction of ET and T at other sites ('spatial generalization'). Results show that under certain conditions single-site temporal prediction of T improves when both data types are combined.

This is an interesting and novel study, but several issues require attention as detailed below, mostly related to methodology (clarifications about data, models, calibration) and readability of the paper (excessive use of machine learning jargon).

R/ Thank you for the careful summary of our study and for recognizing its novelty and relevance. We appreciate that the review identifies methodological clarification and readability as the main areas requiring improvement. We have carefully considered these concerns and will address them through the revisions described below. To maintain consistency with the original numbering, this comment is labeled as Comment 0.

Specific comments:

RC1-1/ 1. The abstract would be easier to follow with less machine learning jargon ("task", "generalization", "supervision", "supervision weight").

R/ We appreciate this suggestion and agree that several machine-learning shorthand terms in the Abstract reduce readability for hydrological readers. **We will simplify the terminology by replacing “within-site temporal generalization” and “cross-site spatial generalization” with “within-site temporal validation” and “cross-site validation,” respectively, and by replacing “ T supervision” and “ T -supervision weight” with direct descriptions of sap-flow-derived T information and its influence during joint training.** The terminology changes related to the 2 validation schemes are discussed further in our responses to Comments 16 and 17, while the terminology related to the use of T information and the loss-weighting parameter is discussed further in Comments 8 and 19, respectively.

Lines 24–27 (Abstract) — modified:

Compared with the baselines of single-task XGBoost and multilayer perceptron models, the multi-task framework primarily improved T estimation, particularly under within-site temporal validation and high evaporative-demand conditions characterized by high vapor pressure deficit, high air temperature, or both.

Lines 27–28 (Abstract) — modified:

Cross-site validation showed stronger site dependence, indicating that the transferability of sparse sap-flow-derived T information remained limited by cross-site process differences and uncertainty in sap-flow-derived T data.

Lines 29–30 (Abstract) — modified:

Input allocation between shared and task-specific branches strongly affected model performance, whereas increasing the influence of T information during joint training alone did not consistently improve performance.

RC1-2/ 2. line 51: "most existing hourly ET modeling studies still treat ET as a single bulk flux, while ET partitioning is often conducted as a separate analysis rather than being integrated into the ET estimation task". I suggest to rephrase and clarify this statement: since process-based ET models exist that separately account for T and E, the statement as currently written is ambiguous and inaccurate.

R/ We agree that the original statement is too broad because “treat ET as a single bulk flux” could be interpreted as implying that existing hourly ET models are generally unable to represent individual ET components, which was not our intended meaning. Process-based hydrological and land-surface models can explicitly represent separate ET components; for example, VIC simulates vegetation transpiration, canopy evaporation, and soil evaporation separately. Recent studies have also extended sub-daily flux estimation beyond total ET . For example, Zhou et al. (2016), Yu et al. (2022), and Nelson et al. (2024) estimate or provide transpiration at sub-daily resolution.

The distinction we intend to make is therefore not whether a model is technically capable of representing individual ET components, but how transpiration observations from an independent measurement system can be incorporated as additional information to constrain the ET – T relationship in hourly joint estimation. In the component-level studies cited above, T is derived from eddy-covariance ET , GPP , or associated carbon–water coupling information rather than from an independent transpiration measurement system such as sap flow.

We will therefore remove the overly broad “single bulk flux” wording and revise the corresponding statements in the

Abstract and Introduction to make this narrower observational distinction explicit.

Lines 17–19 (Abstract) — modified:

However, the same change in total ET may arise from different combinations of vegetation transpiration, soil evaporation, and canopy interception evaporation, making it difficult to interpret transpiration and its sub-daily variability when only total ET is estimated.

Lines 51–53 (Section 1 Introduction) — modified:

Some process-based approaches explicitly represent individual ET components, and recent studies have also extended sub-daily flux analysis beyond total ET by estimating or providing transpiration (Zhou et al., 2016; Yu et al., 2022; Nelson et al., 2024). These developments make clear that the remaining challenge is not whether ET components can be represented. Rather, a distinct question is how transpiration observations from an independent measurement system can be incorporated as additional information to constrain the ET – T relationship in hourly joint estimation. In the component-level studies cited above, T is derived from eddy-covariance ET , GPP , or associated carbon–water coupling information rather than from an independent transpiration measurement system. Thus, the use of independently observed T information to constrain hourly joint ET – T estimation remains comparatively limited.

Lines 54–55 (Section 1 Introduction) — modified:

When only total ET is estimated, it is difficult to distinguish vegetation water use from soil evaporation or to explain the mechanisms underlying sub-daily flux variability (Stoy et al., 2019).

RC1-3/ 3. line 75: "integrating eddy-covariance-based ET references, sap-flow-derived T labels, and GIMMS LAI4g canopy information". What are ET references and T labels?

R/ Here, the 2 terms refer to observation-derived estimates from 2 different measurement systems. **" ET references" denote hourly ET estimates based on FLUXNET2015 eddy-covariance data, whereas " T labels" denote hourly site-scale T estimates derived by upscaling SAPFLUXNET sap-flow measurements.** Here, "label" refers only to the role of these T estimates as target values during supervised model training. Because the original terms do not directly convey the observational origin and physical meaning of these quantities, we will revise the Introduction to describe the 2 estimates and their data sources directly. The definition and interpretation of "reference value" throughout the manuscript are addressed separately in our response to Comment 5.

Lines 74–76 (Section 1 Introduction) — modified:

Here, we develop a physically constrained multi-task learning framework, termed MLF-ETT, for hourly joint ET – T estimation by integrating hourly ET estimates based on FLUXNET2015 eddy-covariance data, site-scale T estimates derived from SAPFLUXNET sap-flow measurements, and GIMMS LAI4g canopy information.

RC1-4/ 4. line 79: "Third, it evaluates how environmental-condition windows, input feature combinations, and T-supervision strength affect model performance under within-site temporal and cross-site validation." Please rewrite in a more understandable manner.

R/ We appreciate this suggestion and agree that the original sentence compresses several different analyses into technical shorthand and is therefore difficult to interpret. **We will therefore rewrite the contribution statement using more**

direct descriptions of the environmental conditions, model inputs, and influence of T information examined in the study.

Specifically, “environmental-condition windows” refers to evaluation subsets defined by particular atmospheric conditions, including high relative humidity, high vapor pressure deficit, high air temperature, and the combined high- VPD -and-high-temperature condition. In the Introduction, we will describe these simply as different environmental conditions. “Input feature combinations” refers to different combinations and allocations of meteorological, radiative, and vegetation variables, which will also be stated explicitly. Finally, “ T -supervision strength” refers to changing the influence of sap-flow-derived T information during joint training through the relative weighting of the T -loss term. Because λ_T and the term “ T -loss weight” have not yet been formally introduced at this point in the Introduction, we will describe this directly as varying the influence of T information during joint training. The corresponding terminology is clarified further in Comments 8 and 19.

The validation terminology will also follow Comments 16 and 17, using “within-site temporal validation” and “cross-site validation.” In addition, the revised contribution statement will retain the internal-relationship and fixed-hyperparameter analyses included in the revised manuscript so that it reflects the full scope of the analysis.

Lines 79–81 (Section 1 Introduction) — modified:

Third, it examines how model performance changes under different environmental conditions, with different combinations of meteorological, radiative, and vegetation inputs, and when the influence of T information during joint training is varied. It further analyzes internal ET – T learning relationships under different environmental conditions and uses fixed-hyperparameter retraining to test whether input-allocation effects persist after controlling for hyperparameter selection, with these analyses conducted under both within-site temporal validation and cross-site validation.

RC1-5/ 5. line 92: "These data were used to derive site-scale T reference values from tree-level measurements". Are 'reference' values assumed to be free of error? This terminology is used throughout the paper, please clarify what is meant by a reference value.

R/ We agree that the term “reference value” could be misunderstood as an error-free observational truth. **In this study, “reference value” denotes an observation-derived ET or T estimate used as a target for model training or as the basis for model evaluation; it does not imply that the estimate is free of error.** For sap-flow-derived T estimates, the species basal-area coverage component of the T -sample weight adjusts their relative contribution according to the degree of stand-level observational support. It does not represent a complete quantification of observation error or uncertainty. The definition of species basal-area coverage is clarified in Comment 6, and its calculation and incorporation into the training loss are described in Comment 14.

We will make 2 related revisions:

- (1) In Section 2.1, we will replace “site-scale T reference values” with “hourly site-scale T estimates” when first describing the values derived from SAPFLUXNET tree-level measurements. This avoids introducing the term “reference value” before its meaning has been defined.
- (2) At the beginning of Section 2.2, we will add an explicit definition of “reference value” and retain “ ET reference values” and “ T reference values” thereafter as concise terms for observation-derived estimates used in model training and evaluation.

Line 92 (Section 2.1 Data sources) — modified:

These data were subsequently used to derive hourly site-scale T estimates, as described in Section 2.2.2.

Before Line 100 (Section 2.2 Data processing) — newly added:

Throughout this manuscript, the term “reference value” denotes an observation-derived ET or T estimate used as a target for model training or as the basis for model evaluation; it does not imply an error-free ground truth.

RC1-6/ 6. line 94: "observation coverage was used for sample weighting during model training". What exactly is observation coverage?

R/ The term “observation coverage” is too vague in the current manuscript. **In our implementation, it specifically denotes the proportion of total stand basal area represented by measured species that provide a non-missing species-level sap-flow-derived T estimate at a given hour.**

Specifically, in our T -upscaling procedure, site-scale T is constructed from species-level sap-flow information together with the basal-area contribution of each species. When some measured species do not provide a valid species-level T estimate at a given hour, the corresponding site-scale T estimate is directly supported by only part of the stand. Accordingly, species basal-area coverage is calculated for each hourly T estimate to quantify its stand-level representation. Poyatos et al. (2021) similarly used the proportion of stand basal area covered by sap-flow measurements to characterize the potential for scaling species-level sap-flow observations to stand-level transpiration. Thus, a higher coverage indicates that the corresponding hourly T estimate is supported by sap-flow information representing a larger fraction of the stand.

This coverage is subsequently used as a relative sample-weighting factor during model training, so that T estimates with greater stand-level observational support contribute more strongly to the T loss, other weighting factors being equal. Coverage therefore characterizes differences in stand-level observational support among T estimates, rather than providing a complete quantification of their observation error or uncertainty. Its exact calculation and incorporation into the training loss are further clarified in our response to Comment 14.

To clarify this definition and its role in the manuscript, we will make the following revisions:

- (1) Revise Section 2.1 to specify the SAPFLUXNET stand-structure information used for T upscaling and remove the ambiguous description of “observation coverage.” The final sentence will incorporate the terminology revision described in Comment 5.
- (2) Add the definition of “species basal-area coverage” in Section 2.2.2 after the T -upscaling description.
- (3) The corresponding equation and accompanying text will be revised in detail in our response to Comment 14.

Lines 91–92 (Section 2.1 Data sources) — modified:

The stand-structure metadata included total stand basal area and the proportion of total stand basal area represented by each measured species, which were subsequently used in the tree-to-site upscaling of transpiration (Poyatos et al., 2021).

Before Line 116 (Section 2.2.2 T reference) — newly added:

For each hourly T estimate, species basal-area coverage was further calculated as the fraction of total stand basal area represented by measured species contributing a valid species-level T estimate at that hour.

RC1-7/ 7. line 95: you should say something about the origin of the GIMMS LAI4g data, e.g. are these satellite-

based estimates? Also, the spatial resolution of these data is about ~10km, which is very coarse. Please clarify the choice of this dataset.

R/ We appreciate this comment and agree that the original manuscript did not sufficiently explain either the origin of GIMMS LAI4g or the rationale for using this relatively coarse *LAI* product.

(1) GIMMS LAI4g is a satellite-derived global *LAI* product. Its primary data source is the PKU GIMMS NDVI record derived from AVHRR observations; Landsat *LAI* samples were used as reference data during product development, and reprocessed MODIS *LAI* was subsequently used to extend the temporal coverage to 2020 (Cao et al., 2023). We will identify its satellite origin explicitly in Section 2.1.

(2) We agree that the $1/12^\circ$ spatial resolution is coarse relative to the spatial scale represented by a flux-tower footprint. In this study, however, *LAI* is not used to resolve fine-scale canopy heterogeneity within an individual tower footprint. **GIMMS LAI4g was selected because it provides a consistent long-term *LAI* record covering all study sites and their different observation periods, and it is used here to represent broad seasonal variation in canopy conditions rather than fine-scale canopy heterogeneity within individual tower footprints.**

The source, temporal resolution, and spatial resolution of *LAI* are also summarized together with those of the other 8 model inputs in our response to Comment 20.

Lines 95–97 (Section 2.1 Data sources) — modified:

Canopy structural information was obtained from GIMMS LAI4g (<https://zenodo.org/records/8281930>; accessed on 20 April 2026), a satellite-derived global leaf area index (*LAI*) product covering vegetated regions from 1982 to 2020 at a semi-monthly temporal resolution and a $1/12^\circ$ spatial resolution (Cao et al., 2023). Although this spatial resolution is coarse relative to individual flux-tower footprints, GIMMS LAI4g provides a consistent long-term *LAI* record across all study sites and observation periods and is used here to characterize broad seasonal variation in canopy conditions rather than fine-scale canopy heterogeneity within individual tower footprints.

RC1-8/ 8. line 118: "The derived *T* reference values were further screened to reduce the influence of sap-flow upscaling bias on *T* supervision." Clarify the term '*T* supervision'.

R/ We agree that the term “*T* supervision” is unnecessarily technical and does not clearly describe how the sap-flow-derived *T* data are used in the model. In the original manuscript, “*T* supervision” refers broadly to the use of sap-flow-derived *T* estimates as target information during model training; that is, samples with available *T* estimates contribute to the *T*-loss term during training.

To avoid conflating several different roles of *T*-related information, we will no longer use “*T* supervision” as a general technical term throughout the manuscript. When referring to the observational data themselves, we will use direct expressions such as “sap-flow-derived *T* estimates” or “*T* reference values.” When describing their role in multi-task learning, we will use expressions such as “*T* information during joint training.” The terms “*T*-labeled samples” and “*T*-labeled sites” will be retained only as concise descriptors of samples or sites for which *T* target values are available.

This terminology is also distinct from the 2 weighting quantities used during training. The *T*-sample weight $w_{T,j}$ controls the relative contribution of individual *T* samples within the *T* loss, whereas λ_T controls the relative weight of the entire *T*-loss term in the joint objective. The latter will therefore be referred to specifically as the “*T*-loss weight,” as described in our response to Comment 19.

Accordingly, the general “ T supervision” wording in the Abstract and Introduction will be replaced with more direct descriptions, consistent with our responses to Comments 1 and 4. In Section 2.2.2, the sentence identified by the reviewer will also be rewritten without using the term “ T supervision.”

Lines 118–120 (Section 2.2.2 T reference) — modified:

The resulting T reference values were then checked against the independent ET references for residual site-scale inconsistency (Ford et al., 2007).

RC1-9/ 9. line 120: "a site-specific multiplicative scaling factor was applied to the T reference series to reduce the proportion of samples with $T > ET$ to 5 %". This pre-processing step potentially undersells the benefits of the combined data approach, since by doing this the single- T data models implicitly benefit from the ET data and the $T < ET$ constraint.

R/ We agree with the reviewer’s observation. **Using ET to assess site-scale consistency during T -reference preprocessing does introduce a limited amount of ET -based consistency information into the processed T references used by the single-task T models.** In this preprocessing sense, the single-task T baselines are therefore not completely independent of ET . However, this is distinct from using ET information within the single-task T model itself: ET was not included as a predictor in either the MLP- T or XGBoost- T model, and no $T \leq ET$ constraint was imposed on predictions from either single-task T model. The single-task T models still used only the meteorological, radiative, and vegetation predictors described in the Methods, with the uniformly processed sap-flow-derived T references serving as training targets.

We also found that the original manuscript did not describe the T -reference correction procedure precisely enough. The default HD correction factor of 1.405 was not applied to all HD data. SAPFLUXNET metadata indicate whether species-specific calibration was reported. For HD data reported as having species-specific calibration, the sap-flow-derived T values were retained without an additional default correction; only HD data without reported species-specific calibration received the default factor of 1.405. This avoids reapplying the generic HD correction to data already reported as species-specifically calibrated.

After this stage-1 treatment, the proportion of positive paired ET – T hours with $T > ET$ was evaluated separately at each site. **A downward stage-2 site-specific multiplier ($c_{\text{site}} \leq 1$) was applied only when this proportion exceeded 5%, and the same multiplier was applied uniformly to the entire T reference series at that site.** Thus, individual hourly T values were not adjusted separately according to the corresponding ET value, and the relative temporal dynamics of the sap-flow-derived T series were retained. The ET reference values were not modified.

Importantly, this procedure did not impose a strict $T \leq ET$ constraint on the single-task T data or predictions. Observations with $T > ET$ remaining after the site-level adjustment were retained rather than removed or forcibly corrected. Their contribution to the T loss was reduced through the anomaly factor $\gamma_j = 0.1$. This is a relative sample-weighting step based on observational support and physical inconsistency, rather than a $T \leq ET$ constraint on the single-task model.

The 5% criterion was used as a pragmatic tolerance for residual site-scale inconsistency rather than as a physical constant or statistically optimized threshold. Among the 4 sites reported as species-specifically calibrated, the proportions of positive paired hours with $T > ET$ after the stage-1 treatment were 0.250% at FR-Fon, 2.222% at CA-TP3, 0.591% at NL-Loo, and 5.192% at CA-TP4. Thus, 3 were already below 5%, while CA-TP4 was only slightly above the threshold and required a small stage-2 multiplier of 0.987.

In the final implementation, FR-Fon, CA-TP3, CA-TP4, and NL-Loo were reported as species-specifically calibrated and therefore received no default 1.405 correction. Only CA-TP4 subsequently required the site-level adjustment, giving a final multiplier of 0.987. US-UMB, FR-Pue, and RU-Fyo were not reported as species-specifically calibrated and first received the default 1.405 correction; their subsequent site-level multipliers were approximately 0.783, 0.543, and 0.830, respectively, giving overall multiplicative factors of approximately 1.100, 0.763, and 1.166.

We therefore agree with the reviewer that the processed training targets used by the single-task T models contain limited ET -based consistency information. Because the same T -reference preprocessing was used for single-task and multi-task models, the single-task T baselines already benefit partially from ET -based consistency at the target-preprocessing stage, and the reported multi-task gains should therefore be interpreted relative to these partially ET -informed baselines.

Lines 116–117 (Section 2.2.2 T reference) — modified:

For HD data not reported as having species-specific calibration in SAPFLUXNET, the empirical default correction factor of 1.405 was applied (Flo et al., 2019); no additional default correction was applied to data reported as having species-specific calibration.

Lines 120–125 (Section 2.2.2 T reference) — modified:

For sites where more than 5% of positive paired ET – T samples had $T > ET$, a downward site-specific multiplier was applied uniformly to the full T reference series to reduce this proportion to approximately 5%. This adjustment did not modify the ET reference values or individually force hourly T values below ET . The remaining $T > ET$ samples were retained.

After Line 125 (Section 2.2.2 T reference) — newly added:

Because this site-level consistency adjustment used the independent ET references and was applied before single-task and multi-task training, the processed T references used by the single-task T baselines also contain limited ET -based consistency information.

RC1-10/ 10. line 131: "we retained only HD sites with at least 1000 aligned T-labeled samples". Why this restriction? How many data points are you throwing away by doing this?

R/ The HD restriction and the 1000-h threshold serve 2 different purposes, so we address them separately.

(1) Why only HD sites were retained. The restriction to heat-dissipation (HD) sites was used primarily to reduce methodological inconsistency among sap-flow measurements. In addition to HD, the matched FLUXNET2015–SAPFLUXNET sites included the compensation heat-pulse method (CHP; Bleby et al., 2004), the heat-pulse T-max method (HPTM; Cohen et al., 1981), and the trunk-sector heat-balance method (TSHB; Čermák et al., 1973). CHP and HPTM are based on heat-pulse principles, whereas TSHB estimates sap flow using a heat-balance principle; these methods therefore differ from the thermal-dissipation principle underlying HD (Granier, 1985). More importantly for our processing, the default correction factor of 1.405 used in this study was developed to address the reported systematic underestimation associated with HD measurements (Flo et al., 2019). CHP, HPTM, and TSHB use different measurement equations and have method-specific biases, so applying this HD-specific correction relationship to those methods would not be methodologically appropriate. **We therefore retained only HD sites to maintain a consistent measurement and correction basis for the sap-flow-derived T estimates.** The original manuscript already states the HD restriction and the use of the 1.405 correction factor in Section 2.2.2, but we will make the rationale for the HD restriction more explicit at the site-selection step in Section 2.3.

(2) Why the 1000-h threshold was used and what it excluded. We rechecked the complete site-screening sequence. There were initially 16 mapped FLUXNET2015–SAPFLUXNET site pairs: FR-Fon, US-MMS, US-UMB, US-UMd, AU-Wom, FR-Pue, GF-Guy, CA-TP3, CA-TP4, CH-Dav, CZ-BK1, FI-Hyy, IT-Ren, NL-Loo, RU-Fyo, and US-Syv. Four site pairs (AU-Wom, GF-Guy, CZ-BK1, and IT-Ren) were excluded at the methodological-screening stage because they used non-HD methods. Among the remaining 12 HD sites, US-UMd did not pass the FLUXNET precipitation quality-control requirement before temporal alignment, while CH-Dav, FI-Hyy, and US-Syv yielded no valid paired ET–T observations after data processing and temporal alignment. This left 8 HD sites with valid paired observations. **The subsequent 1000-h criterion itself excluded only 1 additional site, US-MMS, corresponding to 42 valid paired ET–T hours.** The remaining 7 T-labeled sites were FR-Fon, US-UMB, FR-Pue, CA-TP3, CA-TP4, NL-Loo, and RU-Fyo. The 1000-h criterion was therefore used as a pragmatic minimum sample-size requirement for the 5-block within-site temporal validation and site-level model evaluation, rather than as a physically meaningful threshold. We will clarify both selection criteria in Section 2.3.

Lines 132–133 (Section 2.3 Study sites) — modified:

Among sites with co-located sap-flow observations, we retained only HD sites to maintain a consistent sap-flow measurement and correction basis. We further required at least 1000 aligned T-labeled samples as a pragmatic minimum sample-size criterion for the 5-block within-site temporal validation and site-level evaluation; this threshold was not intended to represent a physically meaningful cutoff.

RC1-11/ 11. consider reversing the order of presentation in the methods section: it seems more natural to first present details about the models used, and then provide details about training and evaluation. For example, table 2 is presented before the different models are discussed.

R/ We appreciate this suggestion and agree that the original Methods sequence was not optimal. In the original manuscript, the training loss, validation schemes, and hyperparameter optimization were introduced before the candidate models were described in detail, and Table 2 was also presented before the corresponding model architectures. This makes the methodological flow less intuitive.

We will therefore reorganize Section 3 so that the modeling components are introduced before the training and evaluation procedures. After the overall framework, we will first present the input features and the single-task and multi-task models, including Shared-Bottom, MMOE, and PLE. The training loss and sample weights, validation schemes, and hyperparameter optimization and model selection will then be presented together, and Table 2 will be moved to follow the model descriptions. Evaluation metrics will remain after the training and validation methods.

Following the additional internal-relationship analysis introduced in response to Referee #2, Comment 2, this analysis will be placed after the evaluation metrics, and the SHAP-based interpretation will be presented subsequently. The revised organization of Section 3 will therefore be:

- 3.1 Framework design
- 3.2 Machine learning models and input features
- 3.3 Model training and validation
- 3.4 Evaluation metrics
- 3.5 Internal relationship analysis
- 3.6 SHAP-based model interpretation

RC1-12/ 12. The "train model" caption in Figure 2 suggests that the test data are part of the training? And by 'valid' you mean 'validation'?"

R/ We appreciate this comment and agree that the current Figure 2 can give the misleading impression that the outer-test data are part of model training. This was a problem with the schematic presentation rather than with the actual implementation. **The held-out outer-test data were kept completely separate from model development: they were not used for model fitting, hyperparameter selection, early stopping, or calculation of the scaling factor.** They were used only after model selection and final refitting to generate the final estimates and calculate test metrics.

The nested validation procedure differed slightly between the 2 validation settings. Under within-site temporal validation, the retained samples at each *T*-labeled site were ordered chronologically and divided into 5 sequential blocks. In each outer split, 1 block was held out as the outer-test set, while the remaining 4 blocks formed the outer-training set. These four blocks were then used for inner cross-validation and model tuning. Thus, the held-out temporal block remained untouched throughout hyperparameter optimization and was used only for final evaluation.

Under cross-site validation, leave-one-site-out (LOSO) validation was conducted across the 25 ENF sites. In each outer split, 1 complete site was held out as the outer-test site and all remaining sites formed the outer-training set. Within this outer-training set, *T*-labeled and *ET*-only sites were shuffled separately and then distributed round-robin across 5 inner folds so that the limited *T*-labeled sites were spread as evenly as possible. Because only 4 ENF sites had *T* labels, some inner-validation folds contained no *T*-labeled samples. For multi-task model tuning, these folds contributed only to the *ET* component of the validation objective; for the single-task *T* model, folds without *T*-labeled validation samples were skipped. This is more precise than the original manuscript wording, which stated that sites with and without *T* labels were assigned to separate inner folds.

Hyperparameter selection was based exclusively on inner-validation performance within the outer-training data. For neural-network models, early stopping was also applied only within the inner folds. After the optimal hyperparameters were selected, the final model was refitted on the complete outer-training set. For neural-network models, the training duration for this final refit was fixed using the median of the best epochs obtained across the inner folds. Only after this refitting step was the held-out outer-test block or site used for final prediction and metric calculation. The 99th-percentile scaling factor used for *ET* and *T* was likewise calculated only from the outer-training *ET* data in each outer split.

The term “valid” in Figure 2 does indeed mean “validation.” **We will therefore replace “Inner-valid data” and “Inner-valid error” with “Inner-validation data” and “Inner-validation error.”** To remove the visual ambiguity concerning the role of the outer-test data, we will also replace the overall heading “Train model” with “Model development and evaluation” and explicitly route the outer-test data only to the final “Outer-test metrics” step. The outer-test pathway will be visually distinguished from the inner model-development workflow so that it is clear that these data do not enter inner training, inner validation, hyperparameter optimization, early stopping, or model fitting. The validation labels in Figure 2 will also be synchronized with Comments 16 and 17.

Lines 204–205 (Section 3.1.2 Validation schemes) — modified:

Within each outer-training set, *T*-labeled and *ET*-only sites were shuffled separately and then allocated round-robin across 5 inner folds to distribute the available *T*-labeled sites across folds as evenly as possible. Because only 4 ENF sites had *T* labels, some inner-validation folds contained no *T*-labeled samples. For multi-task model tuning, these folds contributed only to *ET* validation; for single-task *T* tuning, folds without *T*-labeled validation samples were skipped.

Before Line 219 (Section 3.1.3 Hyperparameter optimization and model selection) — newly added:

The held-out outer-test block or site was not used for model fitting, hyperparameter selection, early stopping, or

calculation of the scaling factor; it was used only to generate final estimates and calculate test metrics after final outer-loop training.

Figure 2 (around Lines 165–172, Section 3.1 Framework design) — modified:

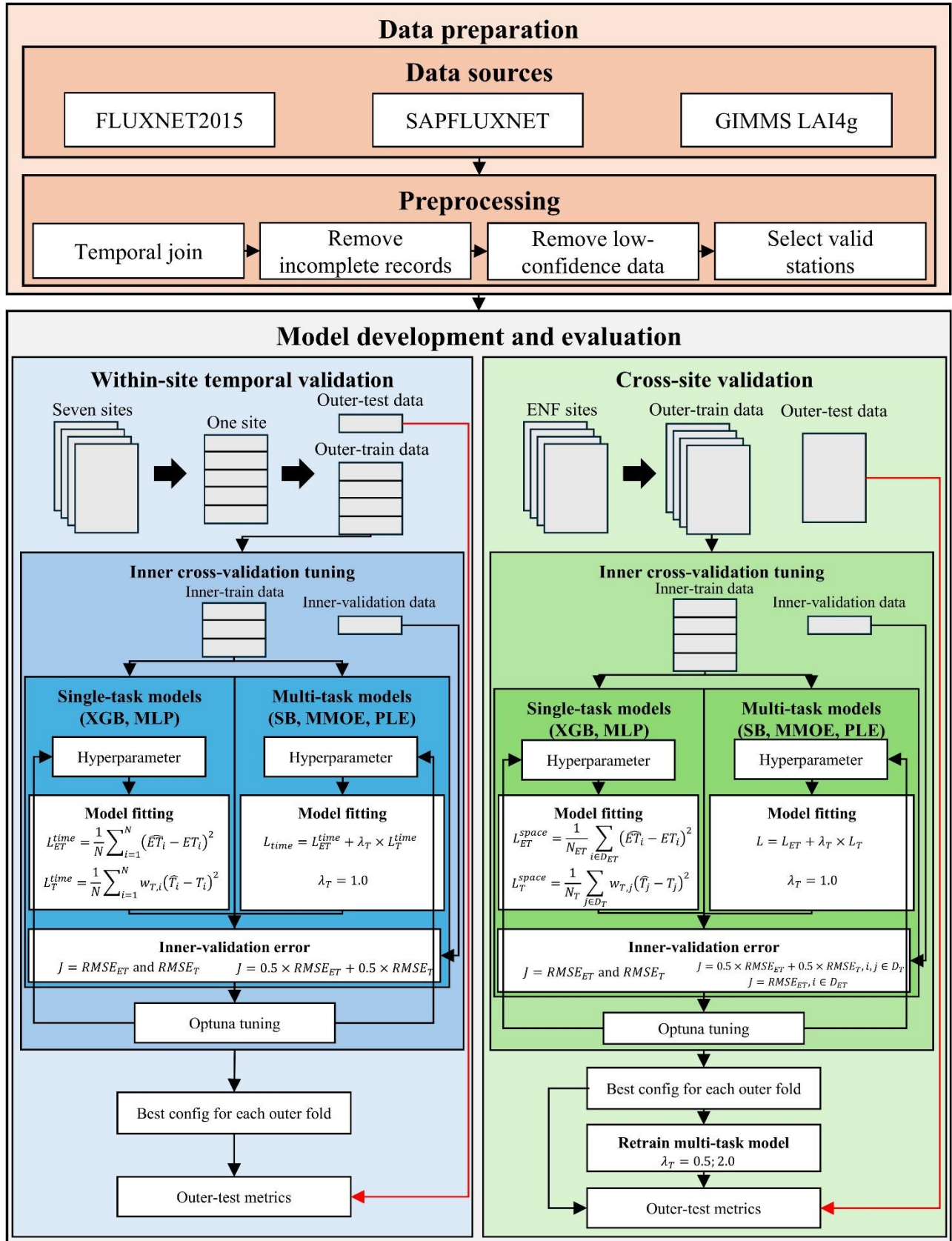


Figure 2. Overall workflow and validation design of the MLF-ETT framework. Note: Within-site temporal validation was evaluated using block-wise validation at T -labeled sites, whereas cross-site validation was evaluated using

leave-one-site-out validation across ENF sites. Hyperparameters were optimized using Optuna within the inner cross-validation loop for each outer split, and final outer-loop training used the selected hyperparameters. The held-out outer-test block or site was used only for final evaluation and did not participate in model fitting, hyperparameter selection, early stopping, or calculation of the scaling factor. Red arrows indicate the pathway of the held-out outer-test data to the final outer-test metrics. L denotes the loss function; λ_T denotes the weight applied to the T loss in the joint training objective; w denotes the sample weight for T -labeled samples; N denotes the number of samples involved in loss calculation under the corresponding scenario; D denotes the set of samples with the corresponding task labels.

RC1-13/ 13. section 3.2.1: the choice of input features requires some further clarification. For example, why is solar elevation angle used as an independent input variable, when net radiation already includes the effect of solar elevation angle? Similarly, there seems to be redundancy in using air temperature, vapor pressure deficit and relative humidity all as independent input variables, since these three quantities are physically related. And two other independent input variables, gross primary production and LAI, are also strongly correlated.

R/ We agree that several of the selected predictors are physically related and can therefore be statistically correlated. **Our intention was not to assume that these variables are mutually independent, but to retain variables that represent complementary aspects of atmospheric forcing, radiation conditions, and vegetation activity at the hourly scale.** The original manuscript did not explain these distinctions sufficiently, particularly the role of solar elevation angle and the different temporal information represented by *GPP* and *LAI*.

For R_n and solar elevation angle (hs), R_n represents the instantaneous net energy available at the surface, whereas hs provides information on solar illumination geometry. Although solar position affects R_n , net radiation alone does not describe how radiation penetrates and is distributed within the vegetation canopy. Solar geometry influences the partitioning of radiation between sunlit and shaded foliage and its vertical distribution within the canopy (Li et al., 2019; Braghiere et al., 2020). Thus, R_n primarily represents net energy availability, whereas hs provides complementary geometric information relevant to canopy radiation distribution.

For T_a , RH , and VPD , we agree that these variables are thermodynamically related rather than statistically independent. However, they describe different aspects of atmospheric conditions. T_a characterizes the thermal state of the atmosphere and the temperature dependence of saturation vapor pressure; RH characterizes atmospheric moisture relative to saturation; and VPD combines temperature-dependent saturation conditions with actual atmospheric moisture to represent evaporative demand relevant to stomatal regulation and transpiration (Monteith, 1995; Grossiord et al., 2020). We therefore retain the 3 variables to represent atmospheric thermal state, relative moisture state, and evaporative demand rather than treating them as independent physical drivers.

GPP and *LAI* also differ in both physical meaning and temporal behavior. As clarified in Comments 7 and 20, *GPP* is obtained from FLUXNET2015 at the site/tower scale and enters the model at hourly resolution, whereas GIMMS LAI4g is provided at semi-monthly resolution and represents more slowly varying canopy structure. *GPP* therefore provides information on rapidly changing vegetation physiological activity and carbon–water coupling relevant to transpiration (Zhou et al., 2016; Nelson et al., 2020), whereas *LAI* provides lower-frequency information on canopy structural state (Cao et al., 2023; Fang et al., 2019). Although the two may covary seasonally, they therefore represent different information in the present hourly framework.

We will revise Section 3.2.1 to make these complementary physical and temporal roles explicit. This clarification does not imply that the predictors are statistically independent; rather, it explains why physically related variables were retained as distinct model inputs.

Lines 234–238 (Section 3.2.1 Input features) — modified:

Although some predictors are physically related, they describe complementary rather than interchangeable aspects of the environmental controls on hourly ET and T . T_a characterizes atmospheric thermal conditions, RH represents atmospheric moisture relative to saturation, and VPD integrates temperature-dependent saturation conditions and atmospheric moisture into a measure of evaporative demand relevant to stomatal regulation and transpiration (Monteith, 1995; Grossiord et al., 2020). R_n represents instantaneous net energy availability, whereas solar elevation angle (hs) characterizes solar radiation geometry relevant to radiation penetration and the distribution of sunlit and shaded foliage within the canopy (Li et al., 2019; Braghieri et al., 2020). GPP provides site-scale, hourly information on vegetation physiological activity associated with rapid transpiration dynamics (Zhou et al., 2016; Nelson et al., 2020), whereas the semi-monthly GIMMS LAI4g provides more slowly varying canopy-structural information (Cao et al., 2023; Fang et al., 2019).

RC1-14/ 14. eq. 4: please clarify how the sap flow observation coverage (T_{coverage}) is computed, what exactly is this number?

R/ The ambiguity in Eq. (4) arises because the original manuscript did not clearly show how the coverage value was calculated or how the reported percentage was converted into the fractional value used in the model. The coverage term in Eq. (4) is the species basal-area coverage defined in Comment 6. For each hourly T estimate, we sum the percentages of total stand basal area represented by measured species that provide a non-missing species-level sap-flow-derived T estimate at that hour, and divide this sum by 100 to obtain a fractional coverage value between 0 and 1.

Specifically, if p_s denotes the percentage of total stand basal area represented by measured species s , and S denotes the set of measured species contributing a non-missing species-level sap-flow-derived T estimate for sample j , the fractional coverage is calculated as $C_{BA,j} = (1/100) \sum_{s \in S_j} p_s$. Thus, $C_{BA,j}$ represents the fractional species basal-area coverage of sample j and ranges from 0 to 1.

We also identified an imprecision in the original notation. The manuscript described T_{coverage} as a percentage, whereas the model implementation used the corresponding fractional value after division by 100. Therefore, the original Eq. (4) did not explicitly show the percentage-to-fraction conversion. This was a notation issue in the manuscript rather than an error in the model implementation.

To make both the calculation of the coverage term and its use in sample weighting explicit, we will make 3 revisions: (1) replace the general notation T_{coverage} with the sample-specific notation $C_{BA,j}$, consistent with the term “species basal-area coverage” defined in Comment 6; (2) revise Eq. (4) to show explicitly both the calculation of $C_{BA,j}$ and its incorporation into the T-sample weight $w_{T,j}$; and (3) revise the accompanying text so that coverage is described as a measure of stand-level observational support rather than as a general indicator of sample reliability. The calculation and interpretation will remain consistent with the explanation provided in Comment 6.

Lines 185–190 (Section 3.1.1 Training loss and sample weights) — modified:

For T -labeled hourly samples, the T loss was further weighted using species basal-area coverage and anomaly flags:

$$C_{BA,j} = (1/100) \sum_{s \in S_j} p_s, w_{T,j} = \gamma_j \times C_{BA,j}, \quad (4)$$

where p_s is the percentage of total stand basal area represented by measured species s , and S_j denotes the set of measured species contributing a non-missing species-level sap-flow-derived T estimate for sample j . Thus, $C_{BA,j}$

represents the fractional species basal-area coverage of sample j and ranges from 0 to 1. The anomaly penalty γ_j was set to 0.1 when $T_j > ET_j$ in the training samples and to 1 otherwise. The resulting $w_{T,j}$ was used as the relative sample weight in the T loss. This weighting reduces the relative influence of T estimates with lower stand-level observational support while retaining their information on T dynamics.

RC1-15/ 15. eq. 5 defines the training objective function, where the first term measures (mis)fit of the ET data, and the second term measures (mis)fit of the T data. We can view this as essentially a weighted nonlinear least squares objective function, where each data point in principle gets a weight that is inversely proportional to its error variance (highly uncertain data points get lower weights). This is indeed how the weight w in eq. 4 is used. However, using that same perspective, the division by N (number of data points) in eq. 3 suggests that simply having more data of a particular type (e.g. ET vs T) reduces the weight and thus the reliability of those data. Please clarify and justify.

R/ We appreciate this question because it highlights an important distinction between sample-count normalization and observation-reliability weighting. **The division by N in Eq. (3) converts the summed squared errors for each task into a mean per-sample loss and does not assign lower reliability to a data type when more observations are available.**

For ET and T , respectively,

$$L_{ET} = \frac{1}{N_{ET}} \sum_{i=1}^{N_{ET}} (\widehat{ET}_i - ET_i)^2, L_T = \frac{1}{N_T} \sum_{j=1}^{N_T} w_{T,j} (\widehat{T}_j - T_j)^2, \quad (3)$$

If the error distribution remains similar as additional observations are added, the number of terms in the summation and the denominator increase together, so the expected mean loss remains on the same scale. Without the $1/N$ normalization, a task with more labeled samples—for example ET under cross-site validation—would produce a larger summed loss simply because it contains more error terms, even if its average per-sample error were identical to that of T . The normalization therefore prevents unequal label availability from mechanically determining the relative magnitude of the 2 task losses. It should not be interpreted as decreasing the reliability of a task as its sample size increases. Additional observations can still improve the statistical stability of model fitting and evaluation; this benefit does not need to be represented by increasing the numerical scale of the task loss.

The T -sample weight $w_{T,j}$ has a different role. As clarified in Comments 6 and 14, it adjusts the relative contribution of individual T samples according to species basal-area coverage and the $T > ET$ anomaly penalty. We agree that this resembles the general idea of giving less influence to less well-supported observations, but $w_{T,j}$ should not be interpreted as a formal inverse-error-variance weight because observation-error variances are not explicitly quantified for the individual sap-flow-derived T estimates.

A third quantity, λ_T , acts at the task level. As clarified in Comment 19, λ_T is the scalar T -loss weight in

$$L = L_{ET} + \lambda_T L_T, \quad (5)$$

and controls the relative contribution of the entire T -loss term to the joint objective.

Thus, the 3 quantities have distinct roles: $1/N$ performs sample-count normalization within each task, $w_{T,j}$ adjusts the relative contribution of individual T samples according to observational support and physical inconsistency, and λ_T controls the relative weighting of the T -loss term during joint training. We will clarify the role of $1/N$ directly after Eq. (3). The broader limitation that the present framework does not explicitly model observation-error variances is addressed in Comment 28.

Lines 184–185 (Section 3.1.1 Training loss and sample weights) — modified:

where N denotes the number of samples with labels for the corresponding task. The factors $1/N_{ET}$ and $1/N_T$ normalize the corresponding summed losses to mean per-sample losses and do not represent observation-reliability weights. The estimated T was calculated as the product of the ET estimate and the estimated T fraction.

RC1-16/ 16. "Within-site temporal generalization": consider replacing/clarifying with temporal prediction at a single site

R/ We agree that the original term “within-site temporal generalization” is more abstract than necessary and does not immediately convey how the validation was performed. **We will therefore replace it with the more direct term “within-site temporal validation.”**

We do not adopt the suggested phrase “temporal prediction at a single site” literally because it could imply a different experimental setting from that used in this study. First, the experiment was not conducted at only 1 site. The same validation procedure was applied separately to all 7 T -labeled sites. Second, the objective was model estimation and validation across held-out time periods within each site rather than time-series forecasting. At each T -labeled site, samples were ordered chronologically and partitioned into 5 sequential blocks, with 1 block held out as the independent test set in each outer split. The detailed construction of these temporal blocks is clarified separately in our response to Comment 18.

We therefore consider “within-site temporal validation” to describe the experimental design more accurately while remaining easier to interpret than the original “within-site temporal generalization.” The term explicitly indicates that validation is conducted along the temporal dimension within each site without implying either forecasting or an experiment conducted at only a single site.

The Abstract will also use “within-site temporal validation,” consistent with our response to Comment 1. We will use the same terminology consistently throughout the remaining manuscript and will clarify the validation procedure explicitly when it is first formally introduced in Section 3.1.2. The corresponding cross-site terminology is addressed separately in our response to Comment 17.

The same terminology will also be updated in relevant figure captions and in figure-panel labels where “Temporal generalization” is currently displayed. These figure revisions change terminology only; the plotted values, model results, and other graphical elements will remain unchanged.

Lines 196–197 (Section 3.1.2 Validation schemes) — modified:

Within-site temporal validation. Within-site temporal validation was used to evaluate model performance across held-out time periods within each T -labeled site.

Corresponding terminology updates: The same terminology will be updated consistently in the remaining occurrences throughout the Introduction, Data, Methods, Results, Discussion, Conclusions, and relevant figure captions. The Abstract will be revised accordingly as described in our response to Comment 1.

Figure revisions: In Figures 2, 8, 9, and 11, figure-panel or workflow labels referring to “Temporal generalization” will be replaced with “Within-site temporal validation.” The corresponding terminology will also be updated in the relevant figure captions, including Figures 4 and 5. These revisions affect terminology only; plotted values and all other graphical elements remain unchanged.

RC1-17/ 17. "Cross-site spatial generalization": consider replacing/clarifying with spatial prediction at multiple sites

R/ We agree with the readability concern raised here, consistent with the related terminology issue discussed in Comment 16. **We will replace “cross-site spatial generalization” with the more direct term “cross-site validation.”**

We do not adopt “spatial prediction at multiple sites” because this phrase could imply spatial interpolation or simultaneous prediction across multiple locations. Our experiment instead uses leave-one-site-out (LOSO) validation, in which 1 complete site is held out for testing in each outer split. Thus, “cross-site validation” more directly describes the actual experimental design.

The detailed LOSO and inner-fold procedure is clarified in our response to Comment 12, where we describe the corresponding revisions to the Section 3.1.2 text and Figure 2. We will use “cross-site validation” consistently throughout the remaining manuscript, figure captions, and figure-panel labels. These terminology changes do not alter the experimental design or reported results.

Line 201 (Section 3.1.2 Validation schemes) — modified:

Cross-site validation. Cross-site validation was used to evaluate model transferability to ENF sites not included in model training.

Corresponding terminology and figure updates: The same terminology will be updated consistently in the remaining occurrences throughout the manuscript and relevant figure captions. Figure-panel or workflow labels referring to “Spatial generalization” will likewise be changed to “Cross-site validation.” The detailed LOSO and inner-fold procedure is clarified separately in our response to Comment 12. These terminology changes do not alter the experimental design or reported results.

RC1-18/ 18. line 195: "samples were ordered chronologically and divided into 5 contiguous temporal blocks". Can you give more details about how long the blocks are and how many data points.

R/ We appreciate this request for clarification. **The 5 temporal blocks were defined by the chronological order and number of retained observations rather than by a fixed calendar duration.** For each of the 7 *T*-labeled sites, retained hourly samples with paired *ET* and *T* reference values and complete predictor data were first ordered by timestamp and then partitioned by sample rank into 5 sequential, non-overlapping blocks containing nearly equal numbers of observations. Here, “sequential” refers to consecutive portions of the ordered retained-sample sequence and does not imply gap-free hourly observations.

Across the 7 sites, 48,641 hourly observations were retained. **Because the partitioning was performed separately at each site, block size depended on the number of retained observations at that site and ranged from 203 to 3,104 observations per block.** For example, NL-Loo contained 1,015 retained observations, so each of its 5 blocks contained 203 observations, whereas FR-Pue contained 15,516 retained observations, resulting in 1 block with 3,104 observations and 4 blocks with 3,103 observations. The total of 48,641 observations is consistent with the *T*-labeled sample counts reported for the 7 sites in Table 1.

The blocks also did not have a common calendar duration because temporal gaps remained after data matching and filtering. Based on the retained timestamps, the shortest elapsed time between the first and last observation of a block was 26.1 days for the third block at NL-Loo, whereas the longest was 1,117.8 days for the second block at RU-Fyo. The

latter block contained only 797 retained hourly observations distributed over that calendar interval, illustrating why elapsed calendar duration should not be interpreted as a continuous sequence of hourly observations.

We will therefore replace the potentially ambiguous phrase “five contiguous temporal blocks” in Section 3.1.2 with a direct description of the actual partitioning procedure. The specific block sizes and calendar-span examples are provided here to answer the reviewer’s question, whereas the manuscript will retain the more concise and reproducible definition based on chronological sample rank. The isolation of each resulting outer-test block from the inner model-development procedure is clarified separately in our response to Comment 12. The terminology “within-site temporal validation” is addressed in Comment 16.

Lines 197–198 (Section 3.1.2 Validation schemes) — modified:

For each T -labeled site, samples were ordered chronologically and partitioned by sample rank into 5 sequential, non-overlapping blocks containing nearly equal numbers of observations.

RC1-19/ 19. line 204: I guess the term "supervision strength" comes from supervised learning in machine learning, but this term may not be clear to the (hydrology) readers of HESS

R/ We agree that “supervision strength” does not clearly communicate the mathematical role of λ_T to hydrological readers. In Eq. (5), λ_T is the scalar coefficient applied to the T -loss term L_T in the joint training objective L . Changing λ_T therefore changes the relative contribution of the entire T -loss term during joint training.

This coefficient is distinct from the T -sample weight $w_{T,j}$. The latter enters L_T at the individual-sample level and controls the relative contribution of different samples with T reference values, whereas λ_T controls the relative contribution of L_T as a whole to the joint objective. **We will therefore use the more explicit term “ T -loss weight” whenever referring specifically to λ_T .**

The more general terminology concerning the use of sap-flow-derived T information during model training is addressed separately in our response to Comment 8. Likewise, in the Abstract and Introduction, where λ_T has not yet been formally defined, we will avoid introducing the technical term “ T -loss weight” prematurely and instead use direct wording describing changes in the influence of T information during joint training, consistent with our responses to Comments 1 and 4.

We will revise the definition of λ_T following Eq. (5) and use “ T -loss weight” consistently after the parameter has been introduced, including in the Methods, Results, Discussion, Conclusions, and the Figure 10 caption. Expressions such as “stronger T supervision” will likewise be replaced by direct descriptions of increasing or decreasing the T -loss weight.

This comment does not require changes to the plotted values or figure-panel elements. The Figure 2 and Figure 10 captions will be revised where they refer to λ_T ; changes to the validation labels within Figure 2 are addressed separately in our responses to Comments 16 and 17.

Line 194 (Section 3.1.1 Training loss and sample weights) — modified:

where λ_T is the scalar weight applied to L_T , thereby controlling the relative weighting of the T -loss term in the joint training objective.

Lines 206–209 (Section 3.1.2 Validation schemes) — modified:

Sensitivity to T -loss weight. The multi-task models were first tuned and trained with $\lambda_T = 1.0$, which was used as the

reference setting. To examine sensitivity to the relative weighting of the T -loss term under cross-site validation, the models were retrained with $\lambda_T = 0.5$ and $\lambda_T = 2.0$ under the same training protocol. Their performance was then compared with that at $\lambda_T = 1.0$ to isolate the effect of the T -loss weight from that of the input-sharing structure.

Corresponding terminology updates: The same terminology will be updated consistently in the remaining occurrences, including the Figure 2 caption and, after λ_T is formally introduced, the Results, Discussion, Conclusions, Figure 10 caption, and other relevant text. These changes are terminology clarifications only and do not alter the reported numerical results or plotted values.

RC1-20/ 20. line 232: please specify the data source for the 9 input variables, including temporal and spatial resolution

R/ We appreciate this suggestion and agree that the original manuscript does not present the data sources, temporal resolutions, and spatial support of the 9 model inputs together in a sufficiently explicit way. At present, this information can only be reconstructed by combining descriptions from the Data and Methods sections, which makes it unnecessarily difficult to identify the origin and effective resolution of each predictor.

T_a, VPD, P_a, WS, RH, R_n, and GPP are obtained from FLUXNET2015. These variables are site-/tower-level observations rather than gridded products, and therefore a fixed pixel-based spatial resolution is not applicable to them. FLUXNET2015 provides these records at half-hourly or hourly temporal resolution depending on the site. In our processing, half-hourly records are aggregated to hourly resolution, whereas records originally provided at hourly resolution remain hourly. The *GPP* input is the FLUXNET2015 gross primary productivity estimate derived using the nighttime respiration partitioning method.

LAI is obtained from GIMMS LAI4g. The product is provided at a semi-monthly temporal resolution and a 1/12° spatial resolution; the semi-monthly *LAI* values are assigned to the corresponding hourly samples during multi-source temporal alignment. As discussed in our response to Comment 7, GIMMS LAI4g is a satellite-derived global *LAI* product, and its relatively coarse spatial resolution is used here to represent broad seasonal variation in canopy conditions rather than fine-scale canopy heterogeneity within an individual flux-tower footprint.

Solar elevation angle (hs) is calculated for each hourly sample from the corresponding site location and timestamp. Its temporal resolution is therefore hourly, and its spatial support is site level, consistent with the site-based FLUXNET2015 inputs.

For clarity, the 9 model inputs can be summarized as follows:

Input	Source	Temporal resolution used in this study	Spatial support / resolution
<i>T_a, VPD, P_a, WS, RH, R_n, and GPP</i>	FLUXNET2015	Hourly, from half-hourly or hourly records	Site/tower level; no fixed gridded resolution
<i>hs</i>	Calculated from site location and timestamp	Hourly	Site level
<i>LAI</i>	GIMMS LAI4g	Semi-monthly values aligned with hourly samples	1/12°

Lines 85–89 (Section 2.1 Data sources) — modified:

FLUXNET2015 (<https://fluxnet.org/data/fluxnet2015-dataset/>; accessed on 20 April 2026) provided site-level half-hourly or hourly eddy-covariance fluxes and environmental variables, including air temperature (T_a), vapor pressure deficit (VPD), air pressure (P_a), wind speed (WS), relative humidity (RH), net radiation (R_n), gross primary productivity (GPP), and latent heat flux (LE) (Pastorello et al., 2020).

Before Line 234 (Section 3.2.1 Input features) — newly added:

Among the 9 input variables, T_a , VPD , P_a , WS , RH , R_n , and GPP were obtained from the site-level FLUXNET2015 records described in Section 2.1, LAI was obtained from GIMMS LAI4g, and solar elevation angle (hs) was calculated from site location and timestamp for each hourly sample, with hourly temporal resolution and site-level spatial support.

RC1-21/ 21. description of the different models: can you include the number of parameters to be estimated in each of these models? It would also be good to indicate which of these modeling approaches have been used before for ET/T modeling and which ones are new to this paper.

R/ We appreciate this suggestion and agree that these 2 aspects should be stated more explicitly:

(1) Model complexity. The neural-network models do not have a single fixed parameter count because their architecture hyperparameters, including network depth, hidden-layer width, and the number of experts for MMOE and PLE, were independently selected within each outer split. **Across the final selected models in the 2 validation schemes, the numbers of trainable parameters ranged from 9,601–177,665 for MLP, 71,938–595,458 for Shared-Bottom, 74,802–2,593,010 for MMOE, and 101,202–2,592,970 for PLE.** These architectures were not explicitly parameter-matched; model complexity was treated as part of the hyperparameter space and was selected using only the inner-validation data.

For XGBoost, a directly comparable neural-network parameter count is not well defined because model complexity is represented by the fitted tree ensemble rather than by a fixed set of dense-network weights and biases. **Across the fitted ET and T models and validation schemes, the selected XGBoost models contained 400–1500 trees, 836–80,406 terminal leaves, and 89–78,906 internal split nodes.** We will therefore report these quantities as measures of XGBoost model complexity rather than treating them as directly equivalent to neural-network parameter counts.

(2) Previous use and novelty of the modeling approaches. XGBoost and MLP are established single-task approaches and have previously been applied to evapotranspiration, latent heat flux, and related environmental prediction problems, as already noted in Section 3.2.2. Shared-Bottom, MMOE, and PLE are likewise established multi-task learning architectures rather than models newly developed in this study. However, in our targeted literature search, we did not identify previous studies applying the canonical Shared-Bottom, MMOE, or PLE architectures to joint $ET-T$ estimation. **We will therefore clarify that the novelty lies in applying and comparing these established multi-task architectures within the physically constrained hourly $ET-T$ framework, rather than in proposing the architectures themselves.**

Following the relocated Table 2 in the reorganized Methods (originally before Line 230) — newly added:

Because architecture hyperparameters were independently selected within each outer split, the number of trainable parameters was not fixed for each neural-network model family. Across the final selected models in the 2 validation schemes, trainable parameter counts ranged from 9,601–177,665 for MLP, 71,938–595,458 for Shared-Bottom, 74,802–2,593,010 for MMOE, and 101,202–2,592,970 for PLE. For XGBoost, a directly comparable neural-network parameter count is not defined because model complexity is determined by the fitted tree ensemble rather than by a fixed set of network weights.

After Line 252 (Section 3.2.2 Single-task models) — newly added:

XGB and MLP are therefore used here as established single-task baselines rather than as newly introduced modeling approaches.

After Line 255 (Section 3.2.3 Multi-task models) — newly added:

Shared-Bottom, MMOE, and PLE are established multi-task learning architectures rather than models newly developed in this study (Caruana, 1993; Ma et al., 2018; Tang et al., 2020). However, in our targeted literature search, we did not identify previous applications of these canonical architectures to joint $ET-T$ estimation; their use here therefore represents an application of established multi-task architectures to the physically constrained $ET-T$ setting.

RC1-22/ 22. Fig. 4: what determines the horizontal position of the site level values (shown as dots)?

R/ The horizontal position of the site-level points does not represent a quantitative variable. The central horizontal position of each group corresponds to the model category. Within each model category, the site-level values are displayed with small fixed horizontal offsets according to the alphabetical order of the site IDs: CA-TP3, CA-TP4, FR-Fon, FR-Pue, NL-Loo, RU-Fyo, and US-UMB. These offsets are used only to separate the points visually.

Thus, the horizontal position of an individual point within a model group reflects only this fixed site ordering; it is not determined by KGE, model performance, or any ranking of the sites. We will add the site ordering to the Figure 4 caption to make this plotting convention explicit.

After Line 350 (Figure 4 caption) — newly added:

Within each model category, site-level values are arranged from left to right in the following fixed site-ID order: CA-TP3, CA-TP4, FR-Fon, FR-Pue, NL-Loo, RU-Fyo, and US-UMB; the horizontal offsets are used only for visual separation and do not represent model performance or site ranking.

RC1-23/ 23. figure 10: consider flipping the colors; I was expecting red = bad, blue = good, but the figure has it the other way around.

R/ We agree that the original color scheme is counterintuitive. **We will therefore reverse the color scale in Figure 10 so that negative ΔKGE values are shown in red and positive ΔKGE values are shown in blue, consistent with the reviewer's suggested convention.** This revision changes only the color mapping; the numerical values, model ordering, and other graphical elements will remain unchanged.

Figure 10 — modified:

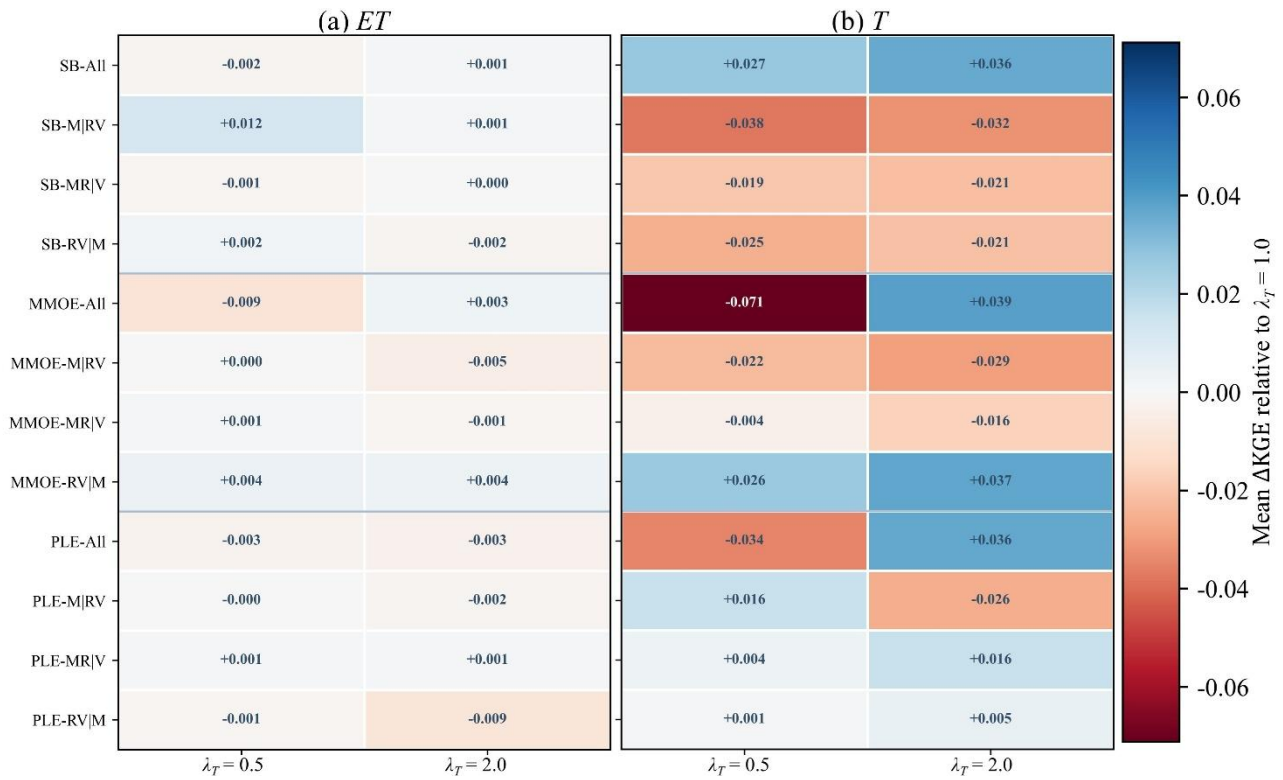


Figure modification: The diverging color scale will be reversed so that negative ΔKGE values are displayed in red and positive ΔKGE values in blue; all plotted values and other graphical elements will remain unchanged.

RC1-24/ 24. section 4.3 - Performance in environmental-condition windows: did you check differences in performance under different vegetation conditions (LAI and gross primary production)?

R/ We appreciate this suggestion and agree that vegetation conditions provide an additional dimension for evaluating when the multi-task framework improves ET and T estimation. **We will therefore add a complementary analysis stratified by LAI and GPP.** To remain consistent with the environmental-condition analysis in Section 4.3, the analysis uses the same selected multi-task configurations and the same relative RMSE gain metric. For each site, Low and High LAI/GPP conditions are defined using the site-specific 20th and 80th percentiles, respectively, of the corresponding full valid hourly record, and these thresholds are then applied to the outer-test samples. These thresholds are used only for post hoc stratification of the evaluation results and do not enter model training, tuning, or selection.

The results show a clearer vegetation-condition dependence for T under within-site temporal validation. **Mean relative T -RMSE gains are +0.0509 (5.1%) and +0.2116 (21.2%) under Low and High LAI , respectively, with positive gains at 3/6 and 6/7 valid sites.** Under Low and High GPP , the corresponding mean gains are +0.0620 (6.2%) and +0.1709 (17.1%), with positive gains at 6/7 and 5/7 sites. Thus, although fewer sites show positive gains under High than Low GPP , the mean T improvement is larger under High GPP . ET mean gains are generally small or negative and do not show a consistent contrast between Low and High LAI or GPP conditions.

The cross-site response is less consistent. High LAI produces a modest positive mean T gain of +0.0417 (4.2%), with positive gains at 3/4 T -labeled test sites, whereas Low LAI shows a mean gain of -0.3106 (-31.1%), with positive gains at only 2/4 sites and pronounced inter-site variability. Mean T gains under Low and High GPP are -0.0480 (-4.8%) and +0.0073 (0.7%), respectively, with only 1/4 sites showing positive gains in either window. Cross-site ET likewise shows no clear and systematic Low–High contrast across the LAI - or GPP -defined windows.

We will report these results in Supplementary Fig. S4 and briefly summarize them in Section 4.3. The denominator of

n^+ denotes the number of sites with valid outer-test samples in the corresponding vegetation-condition window. For example, the Low-*LAI* result under within-site temporal validation is reported as 3/6 rather than 3/7 because US-UMB has no Low-*LAI* outer-test samples within the *ET*–*T* overlap period used for this validation. This does not indicate missing *LAI* or *ET* data at US-UMB; rather, its Low-*LAI* periods do not overlap the retained *T*-labeled evaluation period.

Before Line 419 (Section 4.3 Performance in environmental-condition windows) — newly added:

A complementary analysis further examined model performance under vegetation-condition windows defined by *LAI* and *GPP* (Fig. S4). Under within-site temporal validation, the mean relative *T*-RMSE gains increased from +0.0509 under Low *LAI* to +0.2116 under High *LAI*, corresponding to average site-level *T*-RMSE reductions of 5.1% and 21.2% relative to MLP, with positive gains at 3/6 and 6/7 valid sites, respectively. The corresponding gains under Low and High *GPP* were +0.0620 and +0.1709, equivalent to *T*-RMSE reductions of 6.2% and 17.1%, with positive gains at 6/7 and 5/7 sites. Under cross-site validation, the vegetation-condition response was less consistent: the mean *T* gain was +0.0417 under High *LAI* (4.2% *T*-RMSE reduction; 3/4 sites) but –0.3106 under Low *LAI* (31.1% increase; 2/4 sites), while the Low- and High-*GPP* gains were –0.0480 and +0.0073, respectively. *ET* showed no consistent *LAI*- or *GPP*-dependent improvement in either validation setting.

After Line 503 (Section 5.2 T-estimation gains and their condition dependence) — newly added:

The supplementary *LAI*- and *GPP*-stratified results likewise showed a clearer condition dependence within sites than across sites (Fig. S4). This pattern is consistent with greater site dependence of vegetation-related information during cross-site transfer, but should be interpreted as associative rather than causal.

Supplementary Figure S4 — newly added:

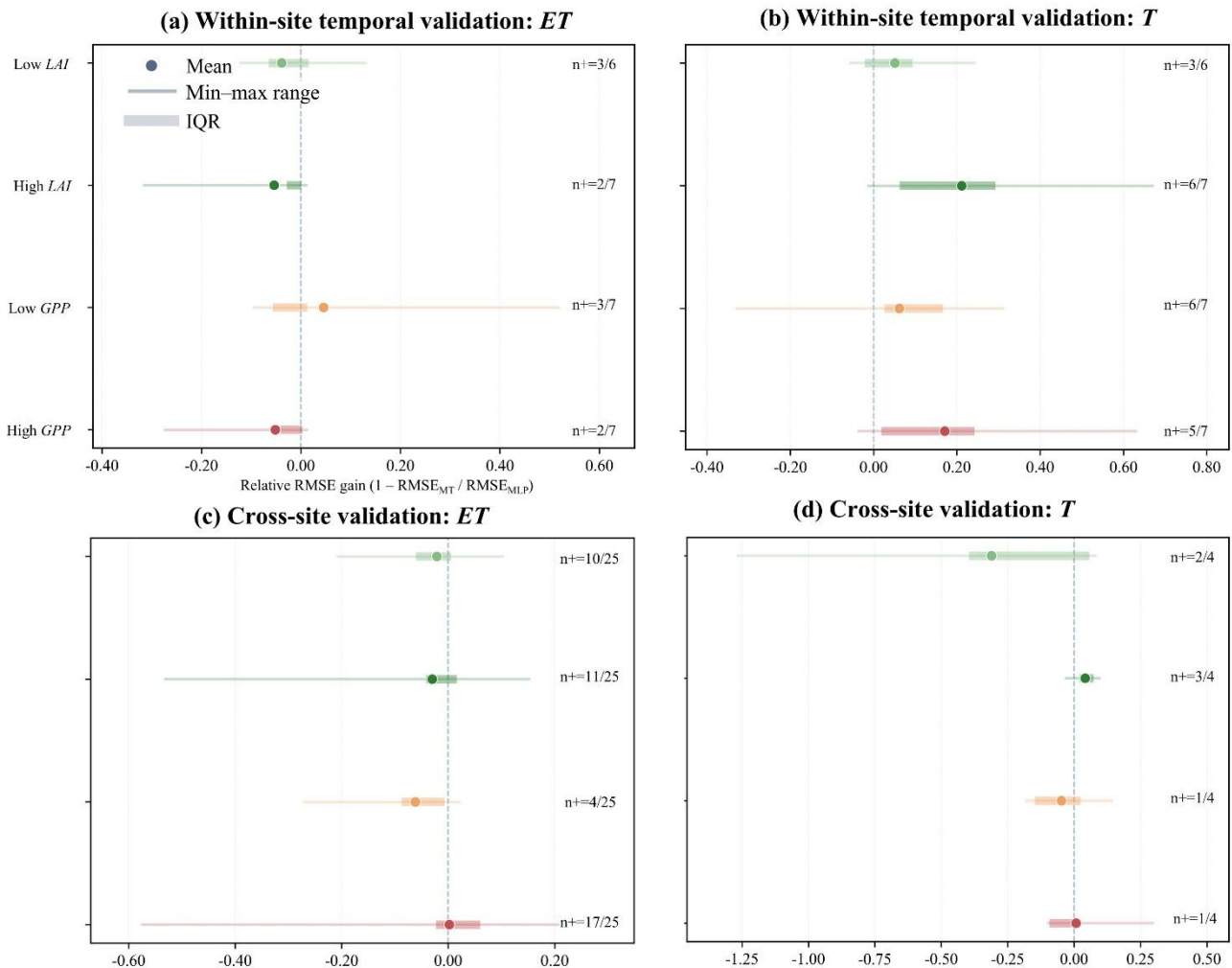


Figure S4. Relative RMSE gain over MLP across vegetation-condition windows. Note: Within-site temporal validation panels use PLE-All, and cross-site validation panels use MMOE-M|RV. Low and High *LAI/GPP* conditions were defined using the site-specific 20th and 80th percentiles, respectively, of the corresponding full valid hourly records; these thresholds were used only for post hoc stratification of outer-test results. Relative RMSE gain was calculated as $1 - \text{RMSE}_{\text{MT}} / \text{RMSE}_{\text{MLP}}$, where MT denotes the selected multi-task model; positive values indicate lower RMSE than MLP. Points, thick bars, and thin lines indicate the mean, interquartile range, and min–max range across sites, respectively. Dashed vertical lines indicate zero gain. The n^+ labels indicate the number of sites with positive relative RMSE gain among sites with valid outer-test samples in the corresponding window.

RC1-25/ 25. line 477: improving the physical consistency of *T* estimates at *ET*-only sites - can you state here explicitly which result (which figure) shows this.

R/ We agree that no figure in the present study directly demonstrates improved *T* accuracy or comparative physical consistency at truly *ET*-only sites. The original wording therefore overstated what can be supported by the present validation.

What the framework provides when local *T* labels are unavailable is a structural property rather than an empirically demonstrated improvement. Because MLF-ETT estimates *ET* and *T/ET* and derives *T* from their product, the estimated *T* remains directly linked to the corresponding *ET* estimate and satisfies the prescribed *ET–T* component relationship. An independently trained *T* model can also extrapolate *T* to a new site, but its *T* prediction is generated independently of the *ET* prediction for the same sample. This difference in model structure does not by itself demonstrate that *T* accuracy or physical consistency is improved at *ET*-only sites.

The cross-site validation also should not be interpreted as direct validation at truly *ET*-only sites. In each outer split, 1 complete site is excluded from model training, providing an independent test of cross-site transferability. However, *T* performance can only be evaluated at held-out sites where sap-flow-derived *T* reference values are available. These sites therefore serve as independent *T*-labeled test sites rather than truly *ET*-only sites. At sites without any independent *T* reference, the framework can still generate a *T* estimate satisfying the prescribed *ET–T* relationship, but its accuracy cannot be directly evaluated in the present study.

We will therefore remove the unsupported phrase “improving the physical consistency of *T* estimates at *ET*-only sites” from Section 5.1 and retain only the supported statement concerning the *ET–T* component relationship. We will also explicitly state in Section 5.1 and in the Conclusions that direct evaluation of *T* accuracy is not possible at truly *ET*-only sites in the present study. The related research question in the Introduction will be narrowed to the transferability of benefits across independent sites when *T* observations are available only at a limited number of sites, consistent with the Introduction restructuring described in our response to Referee #2, Comment 4; therefore, no separate Introduction revision is required under this comment.

Lines 476–478 (Section 5.1 Role of constrained *T* estimates in hourly *ET* modeling) — modified:

In MLF-ETT, the coupling between *ET* and *T/ET* links the estimated *T* directly to the corresponding total *ET*, thereby retaining the prescribed component relationship between *ET* and *T* in joint estimation.

Lines 479–484 (Section 5.1 Role of constrained *T* estimates in hourly *ET* modeling) — modified:

This formulation is particularly useful for *ET*-only sites. Conventional *ET* models provide only total *ET* at these sites, whereas an independently trained *T* model can extrapolate *T* but does not link that estimate to the corresponding *ET* estimate through the prescribed component relationship. By estimating *T/ET* for all samples and multiplying it by estimated *ET*, the proposed framework can generate a *T* component satisfying the prescribed *ET–T* relationship even

when local T labels are unavailable. However, because independent T reference values are unavailable at truly ET -only sites, the accuracy of these T estimates cannot be directly evaluated in the present study.

Lines 561–562 (Section 6 Conclusions) — modified:

Second, direct evaluation of T under cross-site validation was possible only at independent test sites with sap-flow-derived T reference values. At truly ET -only sites, the absence of independent T reference values prevents direct evaluation of T -estimation accuracy.

RC1-26/ 26. The predictions of T and ET are called 'physically constrained', which refers to the $T < ET$ constraint that is enforced (and $ET > 0$). Beyond that, it seems no physical constraints are used, so one wonders how physically realistic the model predictions are. Is there a way to scrutinize the physical plausibility of the predictions (e.g. does the predicted ratio T/ET correlate with LAI)? Maybe the results of the SHAP analysis in Figure 11 can help answer this question, but this part of the analysis is quite short and not very clear.

R/ We agree that the explicit physical constraint in MLF-ETT is the bounded component relationship between ET and T . **The framework estimates ET and T/ET and derives T such that $0 \leq T \leq ET$. This is the specific sense in which we use the term “physically constrained” in the manuscript.** The reviewer’s broader question is therefore whether, beyond satisfying this prescribed relationship, the responses learned from environmental drivers are physically plausible. We agree that this aspect deserves additional scrutiny.

To examine this broader physical plausibility, we will expand the SHAP analysis by adding SHAP dependence for GPP , R_n , T_a , VPD , LAI , and solar elevation (hs) as Supplementary Fig. S5. The horizontal axis of each panel represents the observed value of the corresponding environmental driver, and the vertical axis represents its sample-level SHAP contribution to the estimated ET or T . For PLE configurations in which the same physical variable enters multiple model pathways, its SHAP contributions are summed at the sample level before plotting. Robust LOWESS curves are used to summarize the overall dependence shapes. **These diagnostics do not constitute a complete physical validation of the model, but they provide an additional means of examining whether the learned driver–response relationships are broadly consistent with established controls on ET and T .**

Under within-site temporal validation, ET generally shows increasing SHAP contributions with GPP , R_n , T_a , and VPD , consistent with the roles of vegetation activity, available energy, and atmospheric evaporative demand. For T , GPP shows a clear positive but gradually saturating response. The response to VPD is nonlinear: T contributions generally increase from low to moderate VPD and tend to level off, or in some configurations turn downward, at high VPD . These response shapes are broadly consistent with photosynthesis–stomatal coupling and stomatal regulation as atmospheric demand increases (Medlyn et al., 2011; Grossiord et al., 2020). Solar elevation likewise shows an interpretable nonlinear response, with contributions generally shifting from negative at low hs toward positive values as hs increases and then gradually leveling off at high solar elevations, consistent with the nonlinear influence of solar geometry on canopy illumination (Li et al., 2019; Braghieri et al., 2020).

We also considered the reviewer’s example concerning LAI and T/ET . We do not use a simple LAI – T/ET correlation as a physical-plausibility criterion because a universally monotonic relationship is not expected. LAI can simultaneously affect transpiring leaf area, radiation interception, soil evaporation, and canopy interception, and its net influence on ET partitioning can vary among ecosystems and environmental conditions (Wei et al., 2017). Moreover, Supplementary Fig. S5 shows the SHAP contribution of LAI to ET and T estimates rather than a direct correlation between predicted T/ET and LAI . The LAI dependence is weaker and more configuration-dependent than that of GPP or VPD and does not show a universal monotonic response. We therefore interpret this result as indicating that the learned T response to LAI is not

governed by a simple universal monotonic relationship, rather than as a direct test of a prescribed $LAI-T/ET$ relationship. Under cross-site validation, the dependence shapes are less consistent across configurations and drivers, particularly for T . This pattern is consistent with the weaker and more site-dependent cross-site performance reported in the manuscript and suggests that the learned driver–response relationships are also less stable during cross-site transfer.

We will therefore expand the corresponding Discussion and add Supplementary Fig. S5 to provide a qualitative physical-plausibility diagnostic of the learned environmental responses. We will interpret these results as complementary evidence concerning the plausibility of the learned relationships, rather than as demonstrating complete physical realism of the model.

After Line 532 (Section 5.3) — newly added:

The corresponding SHAP dependence results further show how these key drivers influence model outputs across their observed ranges (Fig. S5). Under within-site temporal validation, ET generally shows increasing contributions from GPP , R_n , T_a , and VPD , whereas T shows a positive but gradually saturating response to GPP and a nonlinear response to VPD that tends to level off, or in some configurations turn downward, under high atmospheric demand. These patterns are broadly consistent with the roles of energy supply, photosynthesis–stomatal coupling, and stomatal regulation in ET and T (Medlyn et al., 2011; Grossiord et al., 2020). LAI shows a weaker and more configuration-dependent relationship with T attribution rather than a universal monotonic response, consistent with its multiple roles in transpiring leaf area, radiation interception, soil evaporation, and canopy interception (Wei et al., 2017). Solar elevation generally shifts from negative contributions at low values toward positive contributions as hs increases, followed by gradual leveling-off at high solar elevations, consistent with the nonlinear influence of solar geometry on canopy illumination (Li et al., 2019; Braghiere et al., 2020). Under cross-site validation, the dependence shapes are less consistent across configurations and drivers, particularly for T , consistent with the stronger site dependence of the cross-site results. These SHAP dependence patterns are interpreted as a qualitative diagnostic of the physical plausibility of the learned driver–response relationships rather than as evidence of complete physical realism.

Supplementary Figure S5 — newly added:

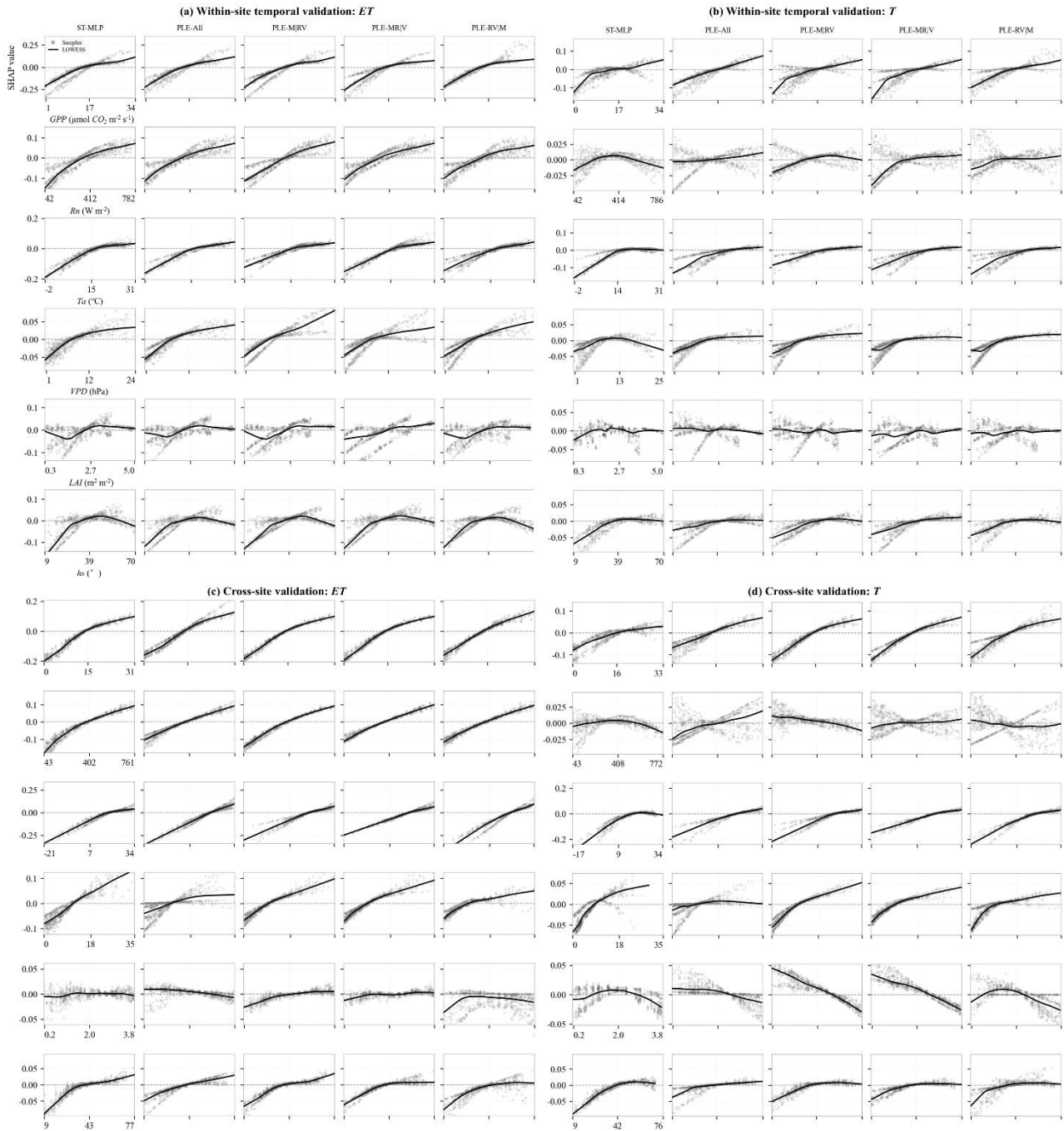


Figure S5. SHAP dependence for key environmental drivers of ET and T estimates. Note: Panels (a, b) show ET and T under within-site temporal validation, and panels (c, d) show ET and T under cross-site validation. Columns compare ST-MLP and PLE input configurations; rows show GPP , R_n , T_a , VPD , LAI , and solar elevation (hs). Grey points show sample-level SHAP values. Black curves show robust LOWESS trends, with each local fit using the nearest 30% of samples followed by 2 residual-based robust reweighting iterations (Cleveland, 1979). For PLE, contributions of the same physical variable entering multiple pathways were summed before plotting. Dashed lines indicate $SHAP = 0$.

RC1-27/ 27. line 549: "The multi-task framework reduced T-estimation errors relative to the single-task model". Not always; this conclusion should be qualified, and would be good to also give some numbers (i.e. by how much were errors reduced).

R/ We agree that the original conclusion was too general because the reduction in T -estimation error was not consistent across all validation settings and environmental conditions. **The clearest and most consistent reductions occurred**

under within-site temporal validation, whereas the cross-site response was more dependent on site and environmental conditions.

Under within-site temporal validation, the selected PLE-All configuration reduced mean T -RMSE by 16.3% across all samples relative to MLP. The reductions increased to 20.2%, 20.7%, and 21.4% under the High VPD , High T_a , and combined High VPD and T_a conditions, respectively, with positive gains at 6/7 sites in each case. By comparison, under cross-site validation, MMOE-M|RV reduced mean T -RMSE by only 1.2% across all samples, while the corresponding reductions under the 3 high evaporative-demand conditions were 9.8%, 8.6%, and 10.2%. The High RH window showed a negative mean gain, with no positive gain at any of the 4 T -labeled test sites. These results show that the multi-task framework did not reduce T -estimation error uniformly across all settings.

We will therefore qualify Conclusion point 2 and add the key quantitative reductions, while distinguishing the stronger within-site temporal improvement from the more site- and condition-dependent cross-site response.

Lines 550-551 (Section 6 Conclusions) — modified:

The multi-task framework showed its clearest reductions in T -estimation error under within-site temporal validation. For the selected PLE-All configuration, mean T -RMSE was reduced by 16.3% across all samples and by 20.2–21.4% under high evaporative-demand conditions relative to MLP; improvements under cross-site validation were more dependent on site and environmental conditions.

RC1-28/ 28. lines 557-560: An important drawback of the methodology is indeed that it does not explicitly account for data errors, although there is some informal weighting of data in the objective functions, but see comment 15.

R/ We agree that the absence of an explicit observation-error model is an important limitation of the present framework. As clarified in Comment 15, the $1/N$ factors in the task losses are used only to normalize summed errors to mean per-sample losses and are not uncertainty weights. Likewise, the T -sample weight $w_{T,j}$, constructed from species basal-area coverage and the anomaly penalty, reflects the relative observational support and physical consistency of individual T estimates rather than a weight derived from explicitly quantified observation-error variance. **The current weighting scheme should therefore not be interpreted as an explicit treatment of data errors.**

This limitation is particularly relevant for the sap-flow-derived T reference values. These estimates may be affected by sample-tree representativeness, sapwood-area estimation, residual calibration uncertainty in HD sap-flow measurements, and uncertainty introduced during tree-to-stand upscaling. The output constraint $0 \leq T \leq ET$ ensures that the predicted T remains within the prescribed physical bounds, but it does not remove uncertainty in the T reference values used for training and evaluation.

A formal treatment of observation errors would require additional information that is not available for the individual ET and T reference values used here, such as independently characterized observation-error variances or probability distributions. If such information were available, it could be incorporated into the loss formulation or a probabilistic training framework so that observations with different uncertainty levels are weighted on an explicit statistical basis.

The present study instead uses ET and T as observation-derived reference values and applies relative T -sample weighting to reduce the influence of samples with lower stand-level observational support or residual physical inconsistency. **We will make this distinction explicit in the limitations section and state that the current framework does not explicitly model or propagate observation-error uncertainty.** The scope of the imposed physical output constraint is discussed further in our response to Comment 26.

Lines 558–560 (Section 6 Conclusions) — modified:

First, the T labels used for model training were derived from upscaled sap-flow measurements and are affected by sample-tree representativeness, sapwood-area estimation, residual calibration uncertainty in HD sap-flow measurements, and tree-to-stand scaling uncertainty.

After Line 561 (Section 6 Conclusions; after the first limitation) — newly added:

The current T -sample weighting represents relative observational support and physical consistency among T estimates but does not explicitly model or propagate observation-error uncertainty.

References

- Bleby, T. M., Burgess, S. S. O., and Adams, M. A.: A validation, comparison and error analysis of two heat-pulse methods for measuring sap flow in *Eucalyptus marginata* saplings, *Functional Plant Biology*, 31, 645-658, <https://doi.org/10.1071/FP04013>, 2004.
- Braghiere, R. K., Quaife, T., Black, E., et al.: Influence of sun zenith angle on canopy clumping and the resulting impacts on photosynthesis, *Agricultural and Forest Meteorology*, 291, 108065, <https://doi.org/10.1016/j.agrformet.2020.108065>, 2020.
- Cao, S., Li, M., Zhu, Z., et al.: Spatiotemporally consistent global dataset of the GIMMS leaf area index (GIMMS LAI4g) from 1982 to 2020, *Earth Syst. Sci. Data*, 15, 4877-4899, <https://doi.org/10.5194/essd-15-4877-2023>, 2023.
- Čermák, J., Deml, M., and Penka, M.: A new method of sap flow rate determination in trees, *Biologia Plantarum*, 15, 171-178, <https://doi.org/10.1007/BF02922390>, 1973.
- Cohen, Y., Fuchs, M., and Green, G. C.: Improvement of the heat pulse method for determining sap flow in trees, *Plant, Cell & Environment*, 4, 391-397, <https://doi.org/10.1111/j.1365-3040.1981.tb02117.x>, 1981.
- Fang, H., Baret, F., Plummer, S., et al.: An Overview of Global Leaf Area Index (LAI): Methods, Products, Validation, and Applications, *Reviews of Geophysics*, 57, 739-799, <https://doi.org/10.1029/2018RG000608>, 2019.
- Flo, V., Martinez-Vilalta, J., Steppe, K., et al.: A synthesis of bias and uncertainty in sap flow methods, *Agricultural and Forest Meteorology*, 271, 362-374, <https://doi.org/10.1016/j.agrformet.2019.03.012>, 2019.
- Granier, A.: Une nouvelle méthode pour la mesure du flux de sève brute dans le tronc des arbres, *Annales des Sciences Forestières*, 42, 193-200, <https://doi.org/10.1051/forest:19850204>, 1985.
- Grossiord, C., Buckley, T. N., Cernusak, L. A., et al.: Plant responses to rising vapor pressure deficit, *New Phytologist*, 226, 1550-1566, <https://doi.org/10.1111/nph.16485>, 2020.
- Li, X., Gentile, P., Lin, C., et al.: A simple and objective method to partition evapotranspiration into transpiration and evaporation at eddy-covariance sites, *Agricultural and Forest Meteorology*, 265, 171-182, <https://doi.org/10.1016/j.agrformet.2018.11.017>, 2019.
- Medlyn, B. E., Duursma, R. A., Eamus, D., et al.: Reconciling the optimal and empirical approaches to modelling stomatal conductance, *Global Change Biology*, 17, 2134-2144, <https://doi.org/10.1111/j.1365-2486.2010.02375.x>, 2011.
- Monteith, J. L.: A reinterpretation of stomatal responses to humidity, *Plant, Cell & Environment*, 18, 357-364, <https://doi.org/10.1111/j.1365-3040.1995.tb00371.x>, 1995.
- Nelson, J. A., Pérez-Priego, O., Zhou, S., et al.: Ecosystem transpiration and evaporation: Insights from three water flux partitioning methods across FLUXNET sites, *Global Change Biology*, 26, 6916-6930, <https://doi.org/10.1111/gcb.15314>, 2020.

- Nelson, J. A., Walther, S., Gans, F., et al.: X-BASE: the first terrestrial carbon and water flux products from an extended data-driven scaling framework, FLUXCOM-X, Biogeosciences, 21, 5079-5115, <https://doi.org/10.5194/bg-21-5079-2024>, 2024.
- Poyatos, R., Granda, V., Flo, V., et al.: Global transpiration data from sap flow measurements: the SAPFLUXNET database, Earth Syst. Sci. Data, 13, 2607-2649, <https://doi.org/10.5194/essd-13-2607-2021>, 2021.
- Wei, Z., Yoshimura, K., Wang, L., et al.: Revisiting the contribution of transpiration to global terrestrial evapotranspiration, Geophysical Research Letters, 44, 2792-2801, <https://doi.org/10.1002/2016GL072235>, 2017.
- Yu, L., Zhou, S., Zhao, X., et al.: Evapotranspiration Partitioning Based on Leaf and Ecosystem Water Use Efficiency, Water Resources Research, 58, e2021WR030629, <https://doi.org/10.1029/2021WR030629>, 2022.
- Zhou, S., Yu, B., Zhang, Y., et al.: Partitioning evapotranspiration based on the concept of underlying water use efficiency, Water Resources Research, 52, 1160-1175, <https://doi.org/10.1002/2015WR017766>, 2016.

Finally, we would like to once again thank the Editor and all the Reviewers for your thorough review and support of our paper. If you have any questions, suggestions, or discussions, please feel free to contact us.