

Response to comments on Manuscript egusphere-2026-2687 (RC2)

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-RC2>

Reply on RC2:

Many thanks for taking the time and effort to review our paper. All comments from Referee #2 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:

RC2-0/ This study proposes a physically constrained multi-task learning framework (MLF-ETT) for hourly joint estimation of evapotranspiration (*ET*) and transpiration (*T*). The topic addresses an important challenge in ecohydrology—the difficulty of partitioning *ET* and the sparsity of sap-flow data. The design combining a *T/ET* output branch with the physical constraint ($0 \leq T \leq ET$) is innovative, and the authors have conducted extensive comparative experiments on feature allocation (shared vs. task-specific) and generalization scenarios (temporal and spatial), demonstrating substantial work.

But several aspects require clarification and further analysis before the paper can be accepted. Below are my specific comments.

R/ Thank you for the careful summary of our study and for recognizing the importance, innovation, and scope of the work. We appreciate that the review identifies several aspects requiring further clarification and analysis before publication. We have carefully considered these concerns and will address them through the revisions and additional analyses described below. To maintain consistency with the original numbering, this comment is labeled as Comment 0.

Specific comments:

RC2-1/ 1. The paper claims that “the multi-task framework reduced T -estimation errors relative to the single-task model.” It would be helpful to present the quantified error reductions (e.g., percentage or absolute RMSE improvements) more explicitly and consistently across the results section.

R/ We appreciate this suggestion and agree that the magnitude of the T -error reductions should be expressed more explicitly and consistently. The original Results already report relative RMSE gains in Section 4.3, but these values are presented as decimal fractions (e.g., +0.1632), which do not immediately communicate the corresponding percentage reduction in RMSE to the reader.

Because the relative RMSE gain in Fig. 8 is defined as $1 - \text{RMSE}_{\text{MT}}/\text{RMSE}_{\text{MLP}}$, these values can be directly interpreted as proportional reductions in RMSE relative to MLP. **We will therefore retain the original relative RMSE gain values so that the text remains directly consistent with Fig. 8, while additionally reporting their corresponding percentage reductions.**

Under within-site temporal validation, the selected PLE-All configuration had a mean relative T -RMSE gain of +0.1632 in the all-sample window, corresponding to an average site-level T -RMSE reduction of 16.3% relative to MLP. In the High VPD , High T_a , and combined High VPD and T_a windows, the mean relative T -RMSE gains were +0.2021, +0.2068, and +0.2135, corresponding to average site-level T -RMSE reductions of 20.2%, 20.7%, and 21.4%, respectively. **Under cross-site validation, the selected MMOE-M|RV configuration had a substantially smaller mean relative T -RMSE gain of +0.0115 in the all-sample window, corresponding to an average reduction of 1.2%, while the corresponding gains in the 3 high evaporative-demand windows were +0.0978, +0.0855, and +0.1021, equivalent to reductions of 9.8%, 8.6%, and 10.2%, respectively.**

Accordingly, we will revise Section 4.3 to report the relative RMSE gain together with its corresponding percentage reduction across the 2 validation settings. We will also summarize the key quantitative values in Conclusion point 2, while distinguishing the stronger within-site temporal improvement from the more site- and condition-dependent cross-site response.

Lines 397–400 (Section 4.3 Performance in environmental-condition windows) — modified:

In the all-sample window, the mean relative RMSE gain for T was +0.1632, corresponding to an average site-level T -RMSE reduction of 16.3% relative to MLP, with positive gains at 6/7 sites. In the High VPD , High T_a , and High VPD and T_a windows, the mean gains increased to +0.2021, +0.2068, and +0.2135, respectively, corresponding to average site-level T -RMSE reductions of 20.2%, 20.7%, and 21.4%, with positive gains at 6/7 sites in all 3 windows.

Lines 403–406 (Section 4.3 Performance in environmental-condition windows) — modified:

In the all-sample window, the mean relative RMSE gain for T was +0.0115, corresponding to an average site-level T -RMSE reduction of 1.2% relative to MLP, with positive gains at 2 of 4 sites. The High RH window showed a negative gain, with no positive gains at any of the 4 T -labeled sites. In contrast, the mean T gains in the High VPD , High T_a , and High VPD and T_a windows were +0.0978, +0.0855, and +0.1021, respectively, corresponding to average site-level T -RMSE reductions of 9.8%, 8.6%, and 10.2%, with positive gains at 3/4 sites in all 3 windows.

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.

RC2-2/ 2. The error analysis currently stays at the level of “which model achieves higher accuracy.” The authors should further attribute the estimation errors to underlying physical reasons and explicitly clarify why the multi-task framework improves ET and T simulations—i.e., what specific mechanism or representation gain enables the improvement, rather than only reporting performance differences.

R/ We appreciate this important comment and agree that the original manuscript mainly compared predictive performance, while the internal learning relationships and physical interpretation associated with the performance differences were not sufficiently examined. We first clarify that the multi-task framework did not consistently improve both ET and T ; the clearer benefits were mainly observed for T , particularly under within-site temporal validation and high evaporative-demand conditions.

To investigate the mechanisms associated with these T gains, we added 3 complementary internal diagnostics: (1) gradient alignment to characterize whether ET and T favor compatible updates of shared parameters (Yu et al., 2020); (2) expert-use metrics to examine how shared and task-specific pathways are used in PLE and MMOE; and (3) CKA to quantify similarity between the internal ET and ratio representations (Kornblith et al., 2019). We also added fixed-hyperparameter retraining to separate the effect of input allocation from that of independently selected hyperparameters.

The added analyses indicate that the internal learning relationship associated with T improvement is validation dependent rather than governed by a single representation-gain mechanism. Under within-site High VPD , the T gain was accompanied by greater ET – T gradient alignment and relatively greater use of shared experts by the T -related branch, consistent with stronger shared learning under a comparatively stable site background. In contrast, cross-site T gains occurred with lower ET – T gradient alignment and more selective ET use of the shared expert pool. We interpret this difference as being consistent with 2 factors: the much broader spatial and environmental coverage of ET reference information than T reference information during cross-site training, and weaker transferability of T -related process relationships across sites with different canopy structure, phenology, soil-water background, and energy partitioning.

The fixed-hyperparameter analysis further showed that neither greater sharing nor stronger task differentiation was universally beneficial. The most stable input allocation differed between within-site temporal validation and cross-site validation, indicating that multi-task performance depends on which meteorological, radiative, and vegetation information is shared and which is retained in task-specific pathways, rather than simply on maximizing information sharing.

Accordingly, we will add Sect. 3.5 for the internal-relationship analysis and Sect. 4.5 for the corresponding results, together with new Figs. 10–11. Sections 5.2–5.3 will be expanded to discuss the validation-dependent mechanisms, unequal ET – T reference-data coverage, cross-site process differences, and input-allocation effects. The new findings will also be summarized briefly in the Abstract and Conclusions, while the Introduction will retain the corresponding analysis objectives. The original sensitivity analysis will become Sect. 4.6, and the SHAP-based analysis will be renumbered accordingly; the substantive SHAP revisions are addressed separately in our response to Comment 3.

After Line 28 (Abstract) — newly added:

Further internal diagnostics showed that the relationship between ET – T information sharing and T improvement differed between within-site and cross-site validation, while fixed-hyperparameter experiments showed that the most stable

allocation of shared and task-specific inputs also depended on the validation setting.

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.

After Line 165 (Section 3.1 Framework design) — newly added:

Additional diagnostic analyses examined internal ET – T learning relationships under different environmental conditions and used fixed-hyperparameter retraining to determine whether input-allocation effects persisted after controlling for hyperparameter selection.

Before Line 320 (after Section 3.4 Evaluation metrics) — newly added:

3.5 Internal relationship analysis of multi-task models

Beyond predictive performance, the effect of multi-task learning also depends on how ET - and T -related information is organized in shared and task-specific pathways. To examine these internal relationships, we diagnosed the direction of gradients on shared parameters, expert-use patterns, and task representations. All diagnostics were calculated post hoc using outer-test samples with valid ET and T reference values and valid T sample weights; they were used only for model diagnosis and were not used for parameter updating, hyperparameter selection, or model selection.

First, gradient alignment was used to quantify whether reducing ET and final T estimation errors favored similar adjustments to the shared model parameters. For each outer-test split, the gradients of the ET loss and final T loss were calculated on the same paired ET – T test samples, with the original T sample weights retained in the T loss. For PLE, the shared parameters were the shared-expert parameters; for MMOE, they were the parameters of all shared experts. Gradient alignment was defined as the cosine similarity between the 2 overall gradients:

$$\mathbf{g}_{ET} = \nabla_{\theta_s} L_{ET}, \quad \mathbf{g}_T = \nabla_{\theta_s} L_T, \quad (18)$$

$$A_{\text{grad}} = \frac{\mathbf{g}_{ET} \cdot \mathbf{g}_T}{\|\mathbf{g}_{ET}\|_2 \|\mathbf{g}_T\|_2}, \quad (19)$$

where θ_s denotes the shared model parameters, ∇_{θ_s} denotes the gradient operator with respect to θ_s , and $\mathbf{g}_{ET}$ and $\mathbf{g}_T$ denote the gradients of the ET and T losses with respect to the shared parameters, respectively. Gradient cosine similarity has been used to diagnose synergy or conflict between tasks in multi-task learning (Yu et al., 2020). The metric ranges from -1 to 1 : higher values indicate that reducing ET and T errors favors more similar adjustments to the shared model components, values near 0 indicate weak directional association, and negative values indicate conflicting optimization directions.

For PLE, which contains both shared and task-specific experts, task-specific expert mass was further defined to describe the extent to which a given task branch relies on its own task-specific pathway:

$$M_t = \frac{1}{N} \sum_{i=1}^N \sum_{k \in E_t} w_{t,ik}, \quad (20)$$

where t denotes the task branch, N is the number of analyzed samples, $w_{t,ik}$ is the normalized gate weight assigned by task t to expert k for sample i , and E_t is the set of task-specific experts for that task. The metric ranges from 0 to 1 . Higher values indicate greater reliance on task-specific experts, whereas lower values indicate relatively greater use of shared experts. This metric describes internal information-pathway use rather than predictive accuracy. For the PLE environmental-condition analysis, this quantity was calculated specifically for the ratio branch and is referred to as ratio-

specific mass.

For MMOE, in which all experts are shared by the 2 tasks, gate concentration was used to describe whether a task concentrates its weights on a smaller subset of shared experts. Gate concentration was calculated as the complement of normalized entropy:

$$C_t = \frac{1}{N} \sum_{i=1}^N \left[1 - \frac{-\sum_{k=1}^K w_{t,ik} \log_2 w_{t,ik}}{\log_2 K} \right], \quad (21)$$

where K is the total number of shared experts and the remaining symbols follow the definitions above. The metric ranges from 0 to 1. Values near 0 indicate relatively even weighting across shared experts, whereas higher values indicate more concentrated use of a smaller subset of shared experts and therefore greater selectivity among shared response patterns. For the cross-site environmental-condition analysis, this quantity was calculated for the *ET* branch and is referred to as *ET*-gate concentration.

Finally, centered kernel alignment (CKA) was used to evaluate whether alternative input allocations changed the internal representations formed by the *ET* and ratio branches before their final outputs. CKA provides a measure of representation similarity in neural networks (Kornblith et al., 2019). For each trained model, the last hidden representations before the output layers were extracted from the *ET* branch and the ratio branch using the same paired *ET*-*T* outer-test samples in the same order. After centering the 2 representation matrices, linear CKA was calculated as:

$$CKA = \frac{\|H_{ET} \cdot H_r\|_F^2}{\|H_{ET} \cdot H_{ET}\|_F \|H_r \cdot H_r\|_F}, \quad (22)$$

where H_{ET} and H_r denote the centered hidden-representation matrices of the *ET* and ratio branches, respectively, and F denotes the *Frobenius norm*. CKA ranges from 0 to 1; higher values indicate more similar internal representations before the *ET* and ratio outputs are formed, whereas lower values indicate greater representation differentiation.

To separate the effect of input allocation itself from differences caused by hyperparameter selection, we conducted fixed-hyperparameter retraining experiments. For each outer split, the All, M|RV, MR|V, and RV|M input configurations had each produced 1 previously selected hyperparameter background. The control experiment retrained each of the 4 actual input allocations under each of these 4 backgrounds, yielding a 4×4 paired design. Within the same outer split and hyperparameter background, only the allocation of meteorological (M), radiative (R), and vegetation (V) information between shared and task-specific pathways was changed; the remaining network structure, training hyperparameters, and training strategy were held fixed. Training length followed the final epoch count selected by the original inner validation, with no new hyperparameter search or early stopping, and the corresponding random seed, data split, and target scaling were retained.

Environmental-condition comparisons used all valid paired *ET*-*T* test samples from the same trained model as the all-sample reference. Fixed-hyperparameter comparisons used the All input allocation under the same hyperparameter background as the reference, where All denotes the allocation in which all M, R, and V variables are provided as shared inputs. For within-site temporal validation, paired changes were first summarized across valid outer folds within each site and then across sites; cross-site validation used the independently held-out sites as the statistical units. Overall responses were summarized by site medians and between-site variability by the interquartile range. Environmental-condition responses in Fig. 10 are reported as signed differences on the original metric scale, calculated as the condition metric minus the corresponding all-sample metric.

Before Line 450 (before the original Section 4.5 sensitivity analysis) — newly added:

4.5 Internal relationships of multi-task models under environmental conditions and input allocations

To further explain the changes in *T* performance under different environmental conditions in Sect. 4.3 and test whether the input-allocation effects in Sect. 4.4 persist after controlling for hyperparameters, we analyzed the internal

relationships of the multi-task models. The environmental-condition analysis used the representative configurations selected in Sect. 4.3: PLE-All for within-site temporal validation and MMOE-M|RV for cross-site validation. For PLE, because the architecture contains both shared and task-specific experts and the preceding environmental analysis focused primarily on T performance, we further examined the dependence of the ratio branch on task-specific experts. For MMOE, in which all experts are shared, we examined the concentration of ET weights across shared experts to characterize task selection within the shared expert pool under cross-site validation. Each environmental condition was paired with all valid environmental samples, and overall responses were summarized using site-level medians, with between-site variability represented by the IQR in Fig. 10.

Within-site temporal validation showed that the improvement in T performance under High VPD was accompanied by stronger ET – T information sharing, whereas high temperature alone did not produce the same internal response (Fig. 10a,b). Under High VPD , the median gradient-alignment difference relative to all valid samples was +0.04, while the ratio-specific mass difference was –0.01, indicating more consistent optimization directions for ET and T and relatively greater use of shared experts by the T -related branch. When High VPD and High T_a occurred together, the 2 metrics changed in the same directions, with differences of +0.02 and –0.01, respectively. In contrast, under High RH , gradient alignment decreased by –0.06 while ratio-specific mass increased by +0.01, showing the opposite internal response; both median differences were approximately 0.00 under High T_a alone. Thus, the clearer T gain under High VPD was not limited to lower prediction error but was also associated with stronger ET – T sharing. However, because site-level directions were not fully consistent for most conditions, this pattern should be interpreted as a median response rather than a uniform change across all sites.

Under cross-site validation, the T improvement under high evaporative demand did not repeat the stronger sharing pattern found within sites; instead, it was accompanied by more concentrated ET selection among shared experts (Fig. 10c,d). Under High VPD and combined High VPD and T_a , the median gradient-alignment differences were –0.05 and –0.07, respectively, whereas the corresponding ET -gate concentration differences were +0.03 and +0.03. Under High T_a alone, gradient alignment changed by +0.01 while ET -gate concentration increased by +0.02. The IQR of ET -gate concentration remained above zero under High VPD , High T_a , and the combined condition, indicating that this concentration tendency was relatively consistent across sites. Under High RH , the median gradient-alignment difference was +0.09, but between-site variability was large. Because all experts in MMOE are shared, greater ET -gate concentration represents more selective use of particular shared response patterns rather than stronger ET – T sharing. Overall, High VPD was associated with improved T performance in both validation settings, but the corresponding internal response differed: within sites it was characterized mainly by stronger ET – T sharing, whereas across sites it was characterized by more selective ET use of the shared expert pool.

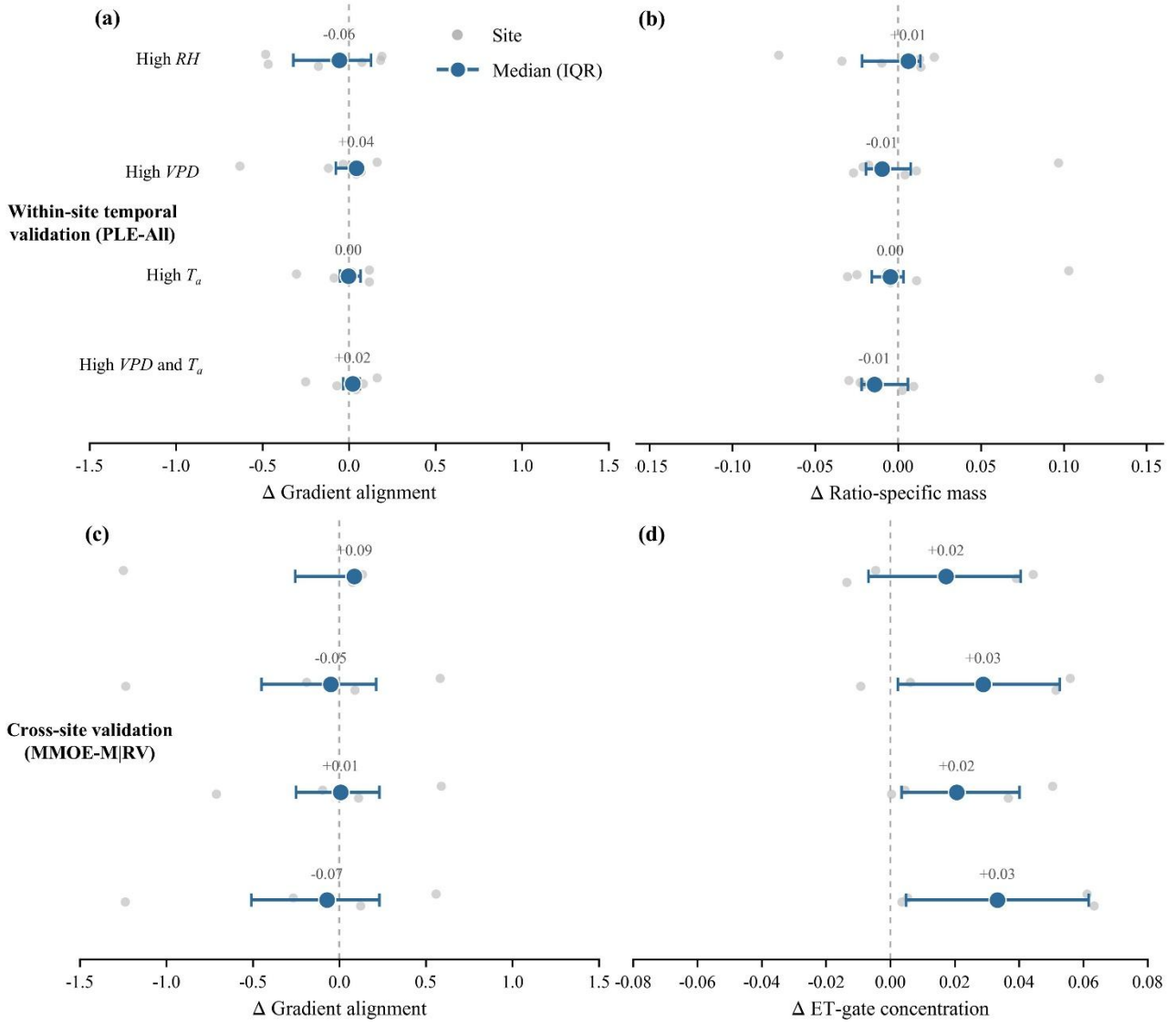


Figure 10. Differences in $ET-T$ task alignment and expert use between selected environmental conditions and all valid test samples. Panels (a,b) show PLE-All within-site results and panels (c,d) MMOE-M|RV cross-site results. For each condition, the plotted difference is the condition metric minus the all-sample metric. Grey dots show site-level paired differences; blue circles and bars show the median and IQR, respectively; the dashed line marks zero.

Fixed-hyperparameter experiments further showed that T -performance differences among input allocations in Sect. 4.4 were not solely caused by independently tuned hyperparameters (Fig. 11). Under within-site temporal validation, MR|V produced the most stable improvement in T performance, whereas RV|M had the strongest effect on internal task differentiation. Across the 4 fixed-hyperparameter backgrounds, the median T -KGE difference for MR|V was +0.0113 and was positive under all 4 settings; RV|M had a median difference of +0.0091 but was positive under only 3 settings, while M|RV changed by -0.0015. Meanwhile, the median gradient-alignment difference for RV|M was -0.0922, with decreases under all 4 settings, and the CKA difference was -0.0571 with the same directional consistency. Thus, moving meteorological variables from shared to task-specific inputs consistently increased differentiation between the 2 internal representations, but this stronger differentiation did not correspond to the most stable T -performance gain; sharing meteorological and radiative information while keeping vegetation information task specific yielded the more stable improvement.

Under cross-site validation, the preferred input allocation changed. M|RV gave the most stable T -performance improvement, whereas RV|M showed a clear adverse effect. The median T -KGE difference for M|RV was +0.0120 and was positive under all 4 fixed-hyperparameter settings; MR|V had a larger median difference of +0.0142 but was positive

under only 3 settings, whereas RV|M had a median difference of -0.0105 and was negative under 3 settings. Gradient alignment was generally lower than All for the 3 alternative allocations, with MR|V and RV|M decreasing under all 4 settings, whereas CKA did not show the consistent decrease observed for RV|M within sites. These results indicate that the most stable organization of shared and task-specific information depended on the validation setting: within-site temporal validation favored sharing meteorological and radiative information while retaining vegetation information as task specific, whereas cross-site validation favored sharing meteorological information only. Together, the 2 settings show that neither greater sharing nor stronger task differentiation alone explains the T -performance response.

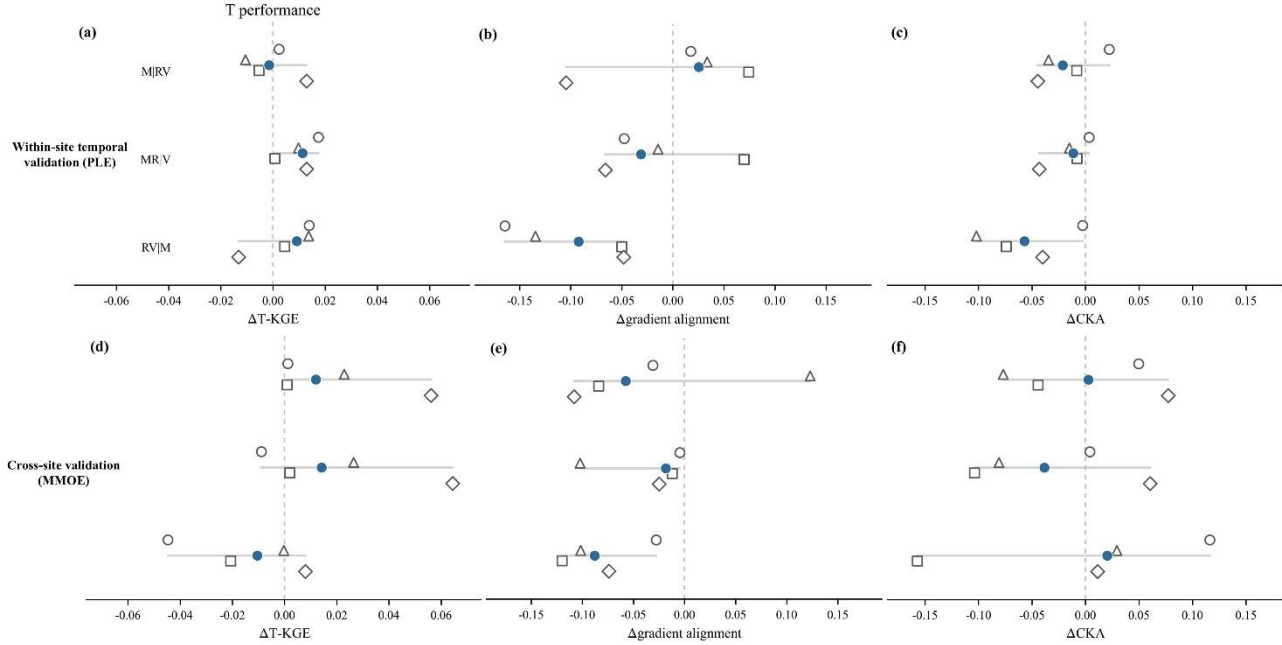


Figure 11. Differences in T performance and internal ET - T relationships between alternative input allocations and All under fixed-hyperparameter settings. Panels (a–c) show PLE within-site results and panels (d–f) MMOE cross-site results; columns show T -KGE, gradient alignment, and CKA differences. Within each fixed-hyperparameter background, each difference is the alternative allocation minus All. Rows represent M|RV, MR|V, and RV|M; open circle, triangle, square, and diamond symbols correspond to hyperparameter backgrounds selected from All, M|RV, MR|V, and RV|M, respectively. Blue circles show medians across the 4 backgrounds, grey lines show their range, and the dashed line marks zero.

Line 450 (original Section 4.5 heading) — modified:

4.6 Sensitivity to T -loss weight

After Line 497 (Section 5.2 T -estimation gains and their condition dependence) — newly added:

An additional source of asymmetry in cross-site training is the unequal coverage of ET and T reference information. ET reference values span nearly all ENF training sites and valid samples, whereas T reference values are restricted to a limited subset of sites and samples. Consequently, ET -related error signals constrain the shared model components across a broader range of sites, environmental states, and samples, while T -related error signals act only within the subset for which valid T reference values are available. Because the ET and T losses are each calculated as mean losses over samples with valid reference values for the corresponding task, the larger ET sample count does not mechanically increase the numerical magnitude of the ET loss through error accumulation. However, this normalization does not compensate for the difference in spatial and environmental coverage of the 2 reference datasets. The shared representations learned during cross-site training are therefore constrained by ET information over a broader range of conditions, while the T constraint is more localized, which may increase the sensitivity of the learned ET - T relationship to site differences.

After Line 502 (Section 5.2; after the sentence discussing High VPD , High T_a , and combined conditions) — newly

added:

The internal-learning-relationship analysis further indicates that T gains under high evaporative demand were associated with different task relationships in the 2 validation settings. Under within-site High VPD , ET and T showed more aligned learning directions on shared parameters, while the T -related ratio branch relied relatively more on shared information. The combined High VPD and T_a condition showed the same directional response, whereas High T_a alone did not show the same clear increase in shared learning. This suggests that the within-site response was more directly associated with atmospheric moisture demand represented by VPD than with high temperature alone. Within a site, canopy structure, vegetation type, and background environmental conditions are comparatively stable, so the effects of atmospheric dryness on transpiration can be repeatedly observed under a more consistent process background and may therefore provide common information that is useful to both ET and T .

Cross-site conditions showed a different relationship. T performance could still improve under High VPD , but the learning directions of ET and T on shared parameters became more differentiated, while ET concentrated more strongly on particular response patterns within the shared MMOE expert pool. Thus, cross-site T gains did not depend on the same increase in ET – T coordination observed within sites and may instead reflect a different organization of task information within the shared representation. Together with the limited coverage of T reference information and differences among sites in canopy structure, phenology, soil-water background, and energy partitioning, this pattern suggests that the same atmospheric-dryness signal does not necessarily correspond to an equally stable T response across sites. T information may therefore retain predictive value under particular environmental conditions while being less able to form a stable shared relationship with ET during cross-site transfer.

After Line 518 (Section 5.3 Input allocation and T -loss weight in multi-task design) — newly added:

The fixed-hyperparameter control analysis further indicates that input allocation itself affects both T performance and the internal ET – T relationship, and that these effects cannot be attributed simply to independent tuning of the different configurations. Under within-site temporal validation, RV|M produced the strongest and most consistent internal task differentiation, whereas the most stable T -performance improvement occurred for MR|V. Stronger differentiation of the internal task representations therefore did not itself result in a more stable T gain; in other words, greater task differentiation is not a sufficient condition for beneficial multi-task learning.

After Line 522 (Section 5.3) — newly added:

The cross-site fixed-hyperparameter analysis further shows that the most stable organization of shared and task-specific information also changes with the validation setting. Within-site temporal validation favored sharing meteorological and radiative information while retaining vegetation information as task specific, whereas the most stable cross-site configuration shared meteorological information only. From within-site to cross-site validation, the range of information that could be shared stably therefore narrowed from meteorological plus radiative information to meteorological information alone. This change is consistent with the process characteristics of the different input groups: meteorological conditions mainly describe external atmospheric forcing experienced by both ET and T , whereas vegetation information represents canopy-structural and physiological controls that can differ substantially among sites. Radiative information can provide a comparatively stable common energy background within a site, but its effects on ET and T are also mediated by canopy structure, radiation penetration, and energy partitioning and may therefore become more site dependent during cross-site transfer.

These results further indicate that the key to multi-task learning is not to share as much information as possible or to maximize task separation, but to identify which drivers can be shared robustly between the 2 tasks and across sites and which information requires task-specific processing.

Meteorological information is particularly informative in this respect. It provides an important external forcing shared by ET and T , but being driven by the same meteorological conditions does not imply that the 2 fluxes respond identically.

T is more directly regulated by stomatal behavior and vegetation physiological activity, whereas total ET also includes soil evaporation and interception evaporation. Multi-task models therefore need both to exploit a common meteorological background and to retain the capacity for ET and T to respond differently to the same forcing.

After Line 526 (Section 5.3; before the concluding sentence on T -loss weight) — newly added:

Given the limited site and sample coverage of T reference information discussed above, increasing the T -loss weight can only change the relative influence of the available T information in the joint objective; it cannot expand the spatial coverage of T information or remove limitations in the cross-site transferability of T -related process relationships.

Lines 552–555 (Section 6 Conclusions, point 3) — modified:

3. Input allocation and T -loss weight affected multi-task performance through different pathways: the former determined how meteorological, radiative, and vegetation information was organized between shared and task-specific branches, whereas the latter changed the contribution of the T -loss term to joint training. Fixed-hyperparameter retraining confirmed that input-allocation effects persisted after controlling for hyperparameter selection; the most stable T improvement occurred with MR|V under within-site temporal validation and M|RV under cross-site validation, showing that neither greater sharing nor stronger task differentiation was universally beneficial.

After Line 555 (Section 6 Conclusions) — newly added:

4. The internal ET – T learning relationships associated with T improvement were validation dependent. Under High VPD , within-site T gains were accompanied by greater ET – T gradient alignment and relatively greater use of shared experts by the T -related branch, whereas cross-site T gains occurred with lower gradient alignment and more selective ET use of the shared expert pool.

Line 528 (Section 5.3) — corresponding figure-number update:

“Figure 11” will be updated to “Figure 13” following the insertion of the new Figs. 10–11 and the corresponding renumbering of the original figures. The original Figure 10 and its corresponding in-text references will be renumbered as Figure 12.

RC2-3/ 3. The SHAP analysis only provides variable importance rankings. It is recommended to reveal the marginal effect shapes of key drivers (e.g., VPD , LAI) on T estimates, for example via partial dependence plots or SHAP dependence plots, to better illustrate how the model responds to these physically meaningful variables.

R/ Thank you for this suggestion. We agree that the original SHAP analysis mainly summarized variable-importance rankings and therefore did not show how model attribution varied across the observed ranges of key environmental drivers. **We have therefore added SHAP dependence plots for GPP , R_n , T_a , VPD , LAI , and solar elevation (hs) as Supplementary Fig. S5.**

In each dependence plot, the original driver value is shown against its sample-level SHAP contribution. For PLE configurations in which the same physical variable enters multiple pathways, its SHAP contributions were summed at the sample level before plotting. A robust LOWESS curve was overlaid only to visualize the overall dependence shape; each local fit used the nearest 30% of samples and was followed by 2 residual-based robust reweighting iterations.

The dependence analysis provides information that was not visible from mean absolute SHAP values alone. Under within-site temporal validation, ET attribution generally became more positive with increasing GPP , R_n , T_a , and VPD . For T , GPP showed a positive but gradually saturating attribution pattern, while the VPD response tended to level off under high atmospheric demand. LAI showed a weaker and less uniformly monotonic relationship with T attribution, and solar elevation showed a nonlinear transition from generally negative attribution at low hs to positive attribution at

higher h_s , followed by a gradual leveling-off. Under cross-site validation, the dependence shapes were less consistent across models and drivers, particularly for T .

These patterns provide a more process-relevant view of how the trained models use key environmental drivers, while remaining an interpretation of model attribution rather than evidence of a direct causal response. We will therefore retain the original mean-absolute-SHAP analysis and supplement it with the dependence plots and corresponding process interpretation in the Discussion. The LOWESS visualization and plotting procedure will be documented in the Supplementary Fig. S5 caption.

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 h_s 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; Braghieri 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:

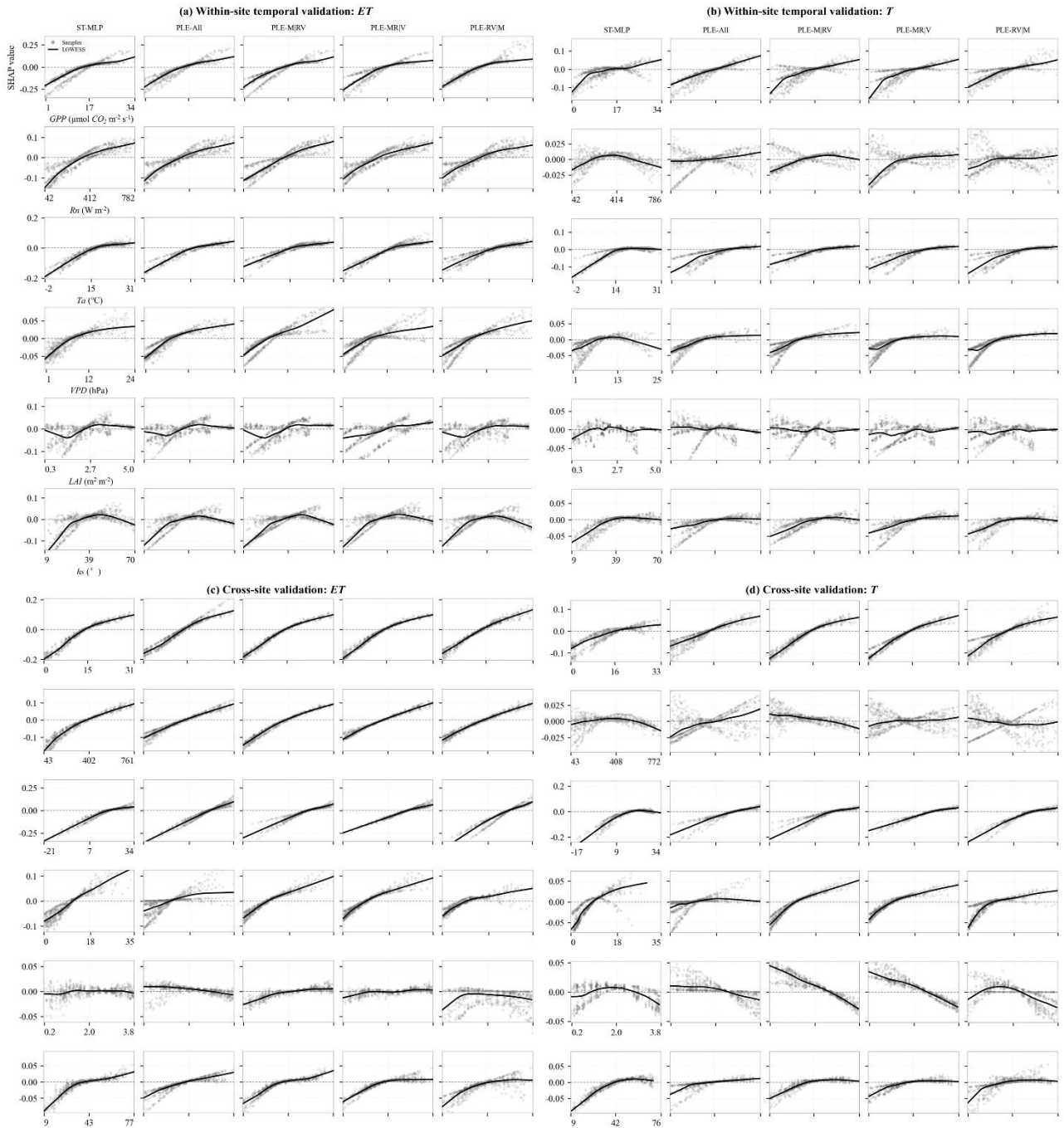


Figure S5. SHAP dependence for key environmental drivers of ET and T estimates. Note: Panels (a, b) show ET and T under within-site temporal validation, and panels (c, d) show ET and T under cross-site validation. Columns compare ST-MLP and PLE input configurations; rows show GPP , R_n , T_a , VPD , LAI , and solar elevation (h_s). 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 $SHAP = 0$.

RC2-4/ 4. The introduction is somewhat diffuse and could be restructured to focus more sharply on the core scientific questions. In particular, the extensive review of general ET estimation methods (physically based and machine learning) and the general applications of multi-task learning are not directly connected to the central theme of this study—i.e., how to incorporate sparse T labels into joint ET – T modeling. A more targeted introduction that builds up to the specific knowledge gaps would strengthen the paper.

R/ We agree that the original Introduction is too diffuse and that the connection between the general literature and the specific problem addressed in this study is not sufficiently direct. **We will therefore reorganize the middle part of the Introduction so that the literature review develops together with the knowledge gaps rather than presenting broad method catalogues before introducing the study motivation.**

(1) For the hourly *ET* literature, we will substantially condense the catalogue of physically based, process-guided, machine-learning, and hybrid methods. The revised paragraph will instead emphasize the distinction that is directly relevant to this study: although some process-based approaches can explicitly represent individual *ET* components and recent studies have also estimated or provided transpiration at sub-daily scales, this does not resolve the separate question of how *T* observations from an independent measurement system can be incorporated as additional information to constrain hourly joint *ET–T* estimation. This distinction will lead directly to the limited use of independently observed *T* information, particularly sap-flow-derived *T*, in hourly *ET–T* modeling.

(2) For the multi-task-learning paragraph, we will remove the catalogue-style discussion of general MTL applications and explain its methodological relevance directly from the unequal availability of *ET* and *T* observations. The revised paragraph will clarify that MTL can accommodate related targets with different numbers of labeled observations, while *ET* and *T* also have a physical component relationship that should be retained in the model formulation. Sadler et al. (2022) will be retained as a hydrological example because it directly illustrates multi-task learning under strongly unequal observation availability between related variables.

Accordingly, the revised Introduction will progress from advances in hourly *ET* estimation, to the remaining gap in using independent *T* observations to constrain hourly *ET–T* estimation, then to the scarcity and current use of sap-flow-derived *T* information, and finally to why multi-task learning is suited to this unequal-observation setting while retaining the physical *ET–T* component relationship. The cross-site research question will also be stated in terms of the transferability of benefits across independent sites rather than direct validation of *T* estimates at *ET*-only sites; the reason for this distinction is discussed specifically in our response to Comment 5.

Lines 40–53 (Section 1 Introduction) — modified:

Hourly actual *ET* estimation has advanced through physically based and process-guided approaches, as well as increasingly diverse machine-learning and hybrid methods, enabling sub-daily flux estimation across a wide range of environmental settings and spatial scales (Su, 2002; Allen et al., 2007; Wang et al., 2023; Nelson et al., 2024; Shi and Cai, 2025; Zhang et al., 2025). 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 54–65 (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). Partitioning *ET* into evaporation and transpiration (*T*) therefore provides a more process-relevant basis for characterizing canopy regulation and vegetation water use (Kool et al., 2014; Schlesinger and Jasechko, 2014; Grossiord et al., 2020). Sap-flow observations provide *T* information from a measurement system independent of eddy-covariance *ET*, and SAPFLUXNET has substantially improved access to such observations by integrating previously fragmented measurements into a standardized global database (Poyatos et

al., 2016, 2021). However, sap-flow observations remain available at relatively few sites, and tree-to-site upscaling introduces additional uncertainty, leaving independently observed T information much sparser than ET observations. Existing studies have mainly used transpiration information for ET partitioning or analyses of T/ET variability (Zhou et al., 2016; Yu et al., 2022; Kar et al., 2025; Xiang et al., 2025), whereas the direct use of independently observed T information as training information in hourly joint $ET-T$ estimation remains limited. This leads to 2 related questions: under what environmental and model-design conditions can limited T information improve T estimation during joint $ET-T$ learning, and how transferable are these benefits across independent sites when T observations are available only at a limited number of sites?

Lines 66–73 (Section 1 Introduction) — modified:

Multi-task learning (MTL) provides a natural framework for this unequal-observation setting because related targets can share information during training while retaining target-specific responses (Zhang and Yang, 2022). MTL does not require all targets to have the same number of labeled samples, allowing each target to use its available observations while shared representations are learned across related targets. This property is directly relevant to hourly $ET-T$ estimation, where ET observations are much more widely available than independent T observations. Sadler et al. (2022), for example, used MTL to jointly model streamflow and water temperature under strongly unequal observation availability, illustrating how information from a sparsely observed hydrological variable can be incorporated alongside a more densely observed related variable. ET and T , however, are not merely statistically related: T is a dynamically varying component of total ET (Wei et al., 2017). An $ET-T$ joint-learning framework therefore needs both to exploit shared information and to retain this physical component relationship in its model formulation.

RC2-5/ 5. In the introduction (page 2, line 64), the authors pose the question: “whether it can support T estimation at sites without local T measurements.” However, the results do not appear to directly address this question with dedicated validation at ET -only sites. The paper should either provide such validation or clearly discuss the limitations and potential pathways to answer this question.

R/ Thank you for raising this point. We agree that the original question, “whether it can support T estimation at sites without local T measurements,” did not clearly distinguish between a structural capability of the framework and an empirically validated performance result.

MLF-ETT can generate T estimates when local T labels are unavailable because it estimates ET and T/ET jointly and derives T from their product, thereby retaining the prescribed $ET-T$ component relationship. However, this structural capability does not by itself demonstrate the accuracy of T estimates at truly ET -only sites. In the present cross-site validation, T performance can be evaluated directly only at independent test sites for which sap-flow-derived T reference values are available. These sites therefore provide an independent test of cross-site transferability, but they are not truly ET -only sites for the purpose of T evaluation.

We will therefore narrow the research question to ask how transferable the benefits of incorporating limited T information are across independent sites when T observations are available only at a limited number of sites, rather than claiming direct validation at ET -only sites. This revised research question is already incorporated into the restructured Introduction described in our response to Comment 4.

We will also revise Section 5.1 to distinguish explicitly between the framework’s structural capability and validated T performance, and will state in the limitations that the absence of independent T reference values prevents direct evaluation of T -estimation accuracy at truly ET -only sites. Future work will identify additional independent transpiration constraints, including sap-flow observations at new test sites, isotope-based partitioning, and

other independent *ET*-partitioning measurements or products, as the basis for such evaluation.

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.

Lines 565–566 (Section 6 Conclusions) — modified:

Future work could incorporate more independent transpiration constraints, including additional sap-flow observations at independent test sites, isotope-based partitioning, and other independent *ET*-partitioning measurements or products, to directly evaluate *T* estimates at sites without local *T* observations and determine how reliably the learned *ET–T* relationships transfer to these sites.

RC2-6/ 6. It is recommended to supplement typical diurnal cycle plots of observed and simulated *ET* and *T* for representative clear-sky days (e.g., a high-*VPD* day and a low-*VPD* day). Such time-series visualizations would greatly help readers intuitively assess the temporal consistency of the estimates, the phase relationships between *ET* and *T*, and the model behavior during key periods (e.g., midday depression), thereby providing process-level evidence for the effectiveness of the physical constraint and the multi-task design.

R/ Thank you for this suggestion. We have added representative clear-sky diurnal trajectories of observed and estimated *ET* and *T* as Supplementary Fig. S3. The figure compares the single-task MLP with the representative multi-task configurations used in Sect. 4.3—PLE-All for within-site temporal validation and MMOE-M|RV for cross-site validation—under All, High *VPD*, Low *VPD*, High *T_a*, and Low *T_a* conditions.

To avoid subjective selection of visually favorable days, representative days were identified using a fully model-independent procedure. We first screened calendar days from the outer-test data. A qualified clear-sky day was required to contain at least 80% of the expected daytime hours, have no precipitation during the calendar day or the preceding 6 h, and satisfy a daily clearness index

$$K_t = \frac{\sum SW_{in}}{\sum SW_{in,POT}},$$

where SW_{in} is the observed incoming shortwave radiation and $SW_{in,POT}$ is the potential incoming shortwave radiation; the summations are taken over the available daytime hours of each calendar day. The K_t threshold was used as a practical literature-supported clear-sky criterion rather than as a universal boundary (Djafer et al., 2017). For within-site temporal

validation, days containing samples from more than 1 outer-test block were additionally excluded.

Among the qualified clear-sky days, the All condition used the complete qualified-day set. High and Low VPD and T_a groups were defined separately for each site using the 80th and 20th percentiles of qualified-day daily mean VPD and T_a , respectively. **Within each condition, we then retained 1 actual calendar day using a medoid procedure rather than selecting a day according to model performance.** Daily mean GPP , R_n , T_a , VPD , LAI , and solar elevation were robust-scaled using the median and interquartile range, and the candidate day with the minimum summed Euclidean distance to all other candidate days in this environmental space was selected as the representative day (Kaufman and Rousseeuw, 1990). The same selected day was used for the single-task and multi-task models within each site and condition, ensuring a direct paired visual comparison. The number of qualified candidate days, n , is shown in each panel; when only 1 qualified day was available, that day was necessarily retained.

The resulting trajectories generally reproduce the broad daytime timing of observed ET and T , including daytime increases, peak or plateau periods, and subsequent declines, although the degree of agreement remains site- and condition-dependent. The plots also make the ET – T phase relationship and individual temporal discrepancies directly visible, while the multi-task estimates retain the prescribed $T \leq ET$ relationship throughout the displayed periods. We therefore interpret Fig. S3 as a time-resolved diagnostic that complements the aggregate performance metrics, rather than as an independent validation of process dynamics or of the physical constraint itself.

Because the screening and representative-day selection are specific to this supplementary visualization, the detailed criteria will be reported in the Supplementary Fig. S3 caption, while the main text will only briefly refer to the resulting temporal patterns.

After Line 412 (Section 4.3 Performance in environmental-condition windows) — newly added:

Representative clear-sky diurnal trajectories further showed that observed and estimated ET and T generally followed similar broad daytime timing across different VPD and T_a conditions (Fig. S3), including daytime increases, peak or plateau periods, and subsequent declines, although the agreement remained site- and condition-dependent. These trajectories provide a time-resolved complement to the window-scale performance metrics.

Supplementary Figure S3 — newly added:

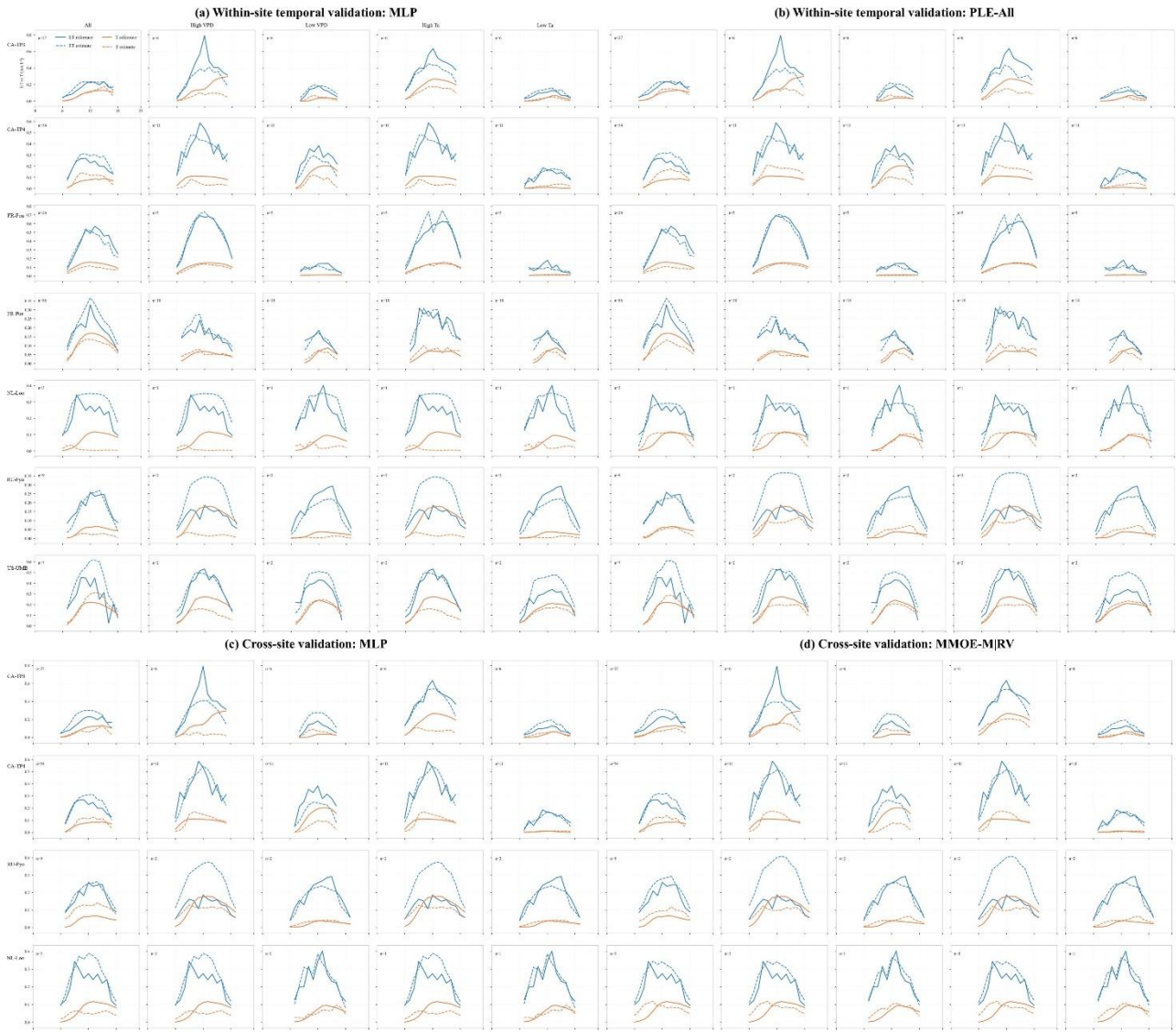


Figure S3. Observed and estimated ET and T on representative clear-sky days. Panels (a,b) show within-site temporal validation using MLP and PLE-All, and panels (c,d) show cross-site validation using MLP and MMOE-M|RV. Columns represent All, High VPD , Low VPD , High T_a , and Low T_a conditions. Qualified days had $\geq 80\%$ daytime-hour coverage, no precipitation during the day or preceding 6 h, and $K_t = \frac{\sum SW_{in}}{\sum SW_{in,POT}} \geq 0.65$ (Djafer et al., 2017). High and Low groups were defined by site-specific 80th and 20th percentiles of daily mean VPD or T_a . One actual representative day per condition was selected as the medoid in robust-scaled daily-mean GPP , R_n , T_a , VPD , LAI , and solar-elevation space (Kaufman and Rousseeuw, 1990). n denotes the number of qualified candidate days.

RC2-7/ 7. Figures 5 and 7 contain too many subplots, which makes them difficult to read. The authors should simplify these figures, for example by showing only representative configurations or sites in the main text, and moving the rest to supplementary materials.

R/ Thank you for this suggestion. We agree that the original Figs. 5 and 7 contained too many subplots, which made the main site-level comparisons difficult to follow. We have therefore simplified both figures in the main text while retaining the complete original comparisons in the Supplementary Material.

In the revised Fig. 5, we retain MLP as the single-task baseline and, for each of the 4 representative T -labeled sites, show only the multi-task configuration with the highest site-level T -KGE: PLE-RV|M for CA-TP3, PLE-M|RV for

CA-TP4, PLE-All for RU-Fyo, and MMOE-RV|M for NL-Loo. This directly preserves the comparison emphasized in Sect. 4.1 between the MLP baseline and the best-performing multi-task configuration for T at each site. The complete original Fig. 5, including XGB, MLP, and all multi-task model–input combinations, will be moved to the Supplementary Material as Fig. S1, with all plotted results retained.

In the revised Fig. 7, we retain only the 2 site–configuration cases explicitly discussed in Sect. 4.2: SB-MR|V at RU-Fyo and PLE-MR|V at NL-Loo. These cases illustrate the site-dependent cross-site response and, in particular, that changes in T -KGE were not necessarily accompanied by corresponding changes in ET -KGE. The complete original Fig. 7 will likewise be moved to the Supplementary Material as Fig. S2.

We will also revise the corresponding main-text references and figure captions so that statements supported by the simplified figures refer to Figs. 5 and 7, whereas statements requiring the complete configuration comparisons refer to Figs. S1 and S2.

Lines 338–339 (Section 4.1 Within-site temporal validation) — modified:

Representative T -labeled sites were further examined using scatter-density plots comparing MLP with the site-specific multi-task configuration yielding the highest T -KGE (Fig. 5); complete model–configuration comparisons are provided in Fig. S1.

Line 355 (Section 4.1 Within-site temporal validation) — modified:

Concurrent ET -KGE and T -KGE gains occurred at some sites (Fig. S1).

Lines 360–364 (Figure 5 caption) — modified:

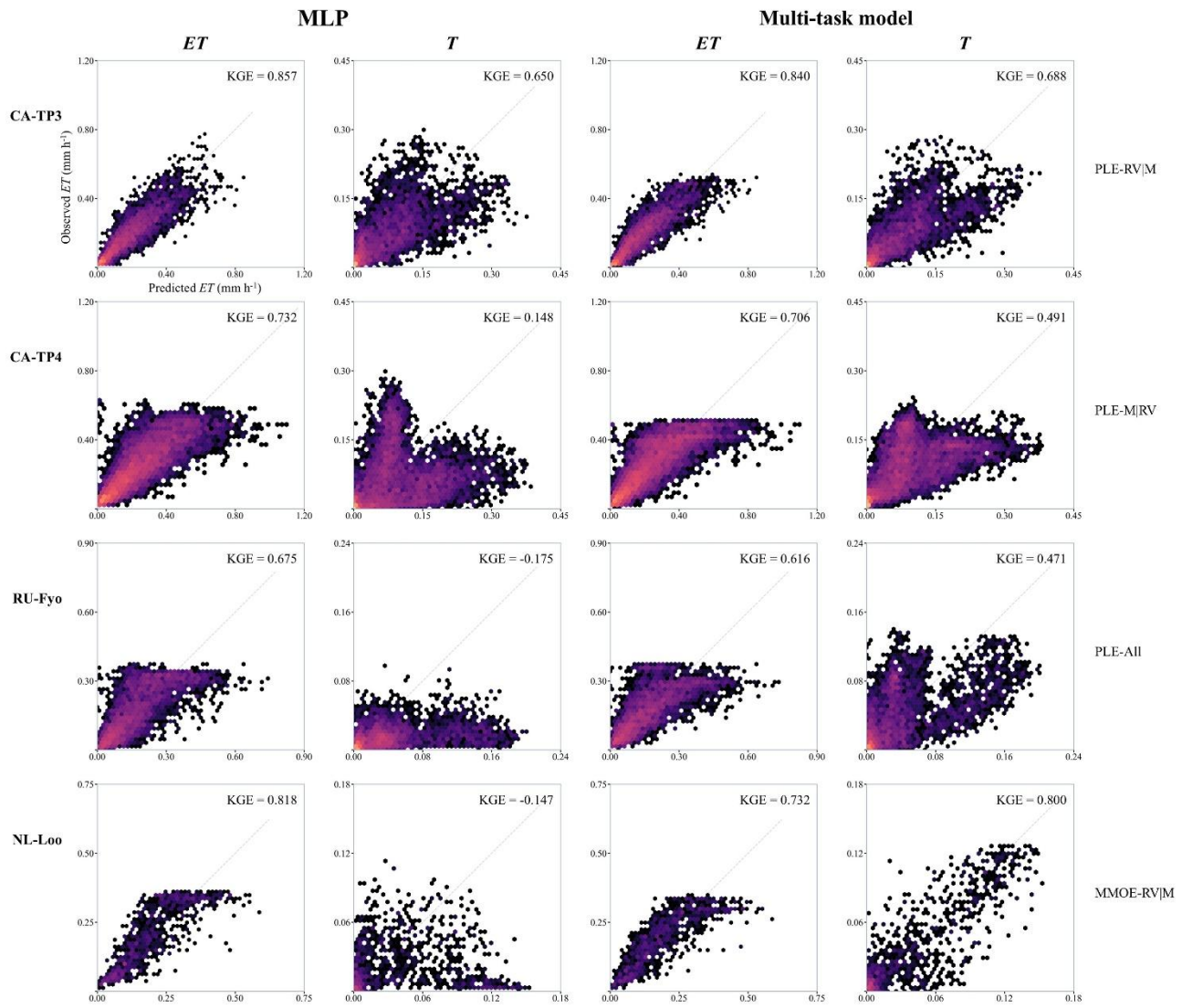


Figure 5. Observed versus estimated ET and T at 4 representative T -labeled sites under within-site temporal validation. MLP is compared with the multi-task configuration yielding the highest T -KGE at each site: PLE-RV|M at CA-TP3, PLE-M|RV at CA-TP4, PLE-All at RU-Fyo, and MMOE-RV|M at NL-Loo. Dashed diagonal lines indicate the 1:1 relationship, and KGE values are shown in each panel.

Lines 367–368 (Section 4.2 Cross-site validation) — modified:

Two representative site–configuration cases were further examined using the scatter-density plots in Fig. 7, while the complete model–configuration comparisons are provided in Fig. S2.

Lines 385–389 (Figure 7 caption) — modified:

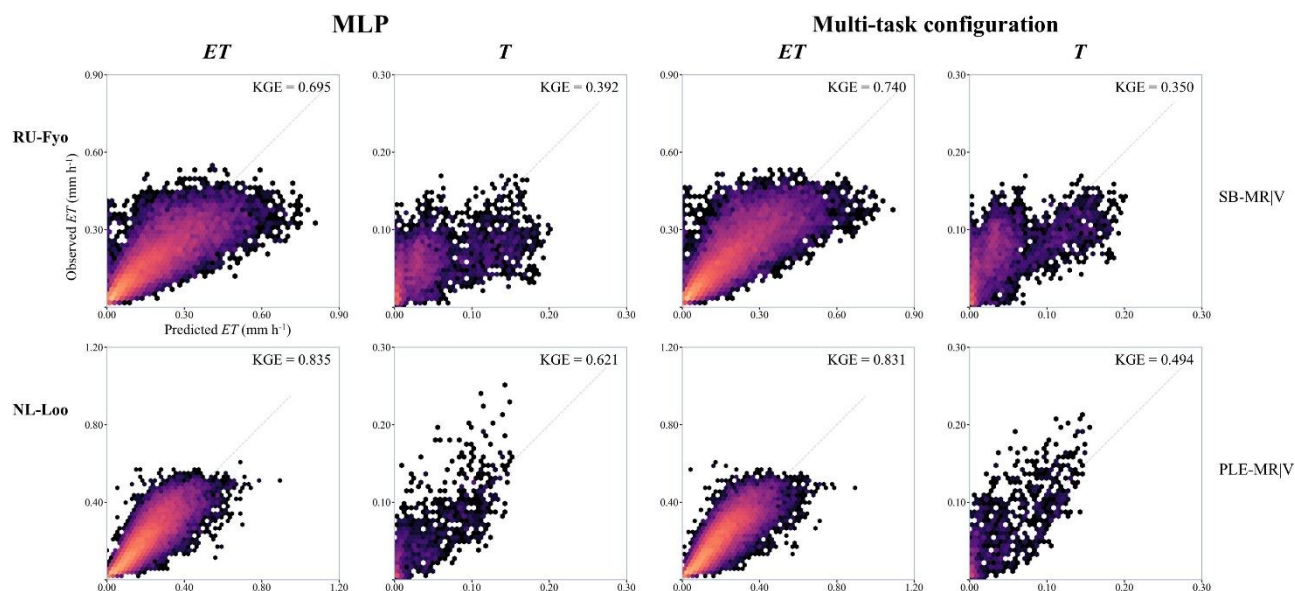


Figure 7. Observed versus estimated ET and T for 2 representative site-configuration cases under cross-site validation. MLP is compared with SB-MR|V at RU-Fyo and PLE-MR|V at NL-Loo. Dashed diagonal lines indicate the 1:1 relationship, and KGE values are shown in each panel.

Supplementary Figure S1 — newly added:

The complete original Fig. 5 will be moved directly to the Supplementary Material with its plotted contents unchanged.

Figure S1. Complete observed-versus-estimated ET and T comparisons at 4 representative T -labeled sites under within-site temporal validation. The left panels show XGB and MLP, whereas the remaining panels show multi-task models with different input feature combinations. In the multi-task panels, M, R, and V denote meteorological, radiative, and vegetation feature groups, respectively; features before and after “|” indicate shared and branch-specific inputs. Rows indicate SB, MMOE, and PLE. Dashed diagonal lines indicate the 1:1 relationship, and color bars indicate sample counts in the hexbin plots.

Supplementary Figure S2 — newly added:

The complete original Fig. 7 will be moved directly to the Supplementary Material with its plotted contents unchanged.

Figure S2. Complete observed-versus-estimated ET and T comparisons at 4 representative T -labeled sites under cross-site validation. The left panels show XGB and MLP, whereas the remaining panels show multi-task models with different input feature combinations. In the multi-task panels, M, R, and V denote meteorological, radiative, and vegetation feature groups, respectively; features before and after “|” indicate shared and branch-specific inputs. Rows indicate SB, MMOE, and PLE. Dashed diagonal lines indicate the 1:1 relationship, and color bars indicate sample counts in the hexbin plots.

References

- Djafer, D., Irbah, A., and Zaiani, M.: Identification of clear days from solar irradiance observations using a new method based on the wavelet transform, *Renewable Energy*, 101, 347-355, <https://doi.org/10.1016/j.renene.2016.08.038>, 2017.
- Kaufman, L. and Rousseeuw, P. J.: *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, <https://doi.org/10.1002/9780470316801>, 1990.

Kornblith, S., Norouzi, M., Lee, H., et al.: Similarity of Neural Network Representations Revisited, Proceedings of the 36th International Conference on Machine Learning, 97, 3519-3529, <https://proceedings.mlr.press/v97/kornblith19a.html>, 2019.

Sadler, J. M., Appling, A. P., Read, J. S., et al.: Multi-Task Deep Learning of Daily Streamflow and Water Temperature, Water Resources Research, 58, e2021WR030138, <https://doi.org/10.1029/2021WR030138>, 2022.

Yu, T., Kumar, S., Gupta, A., et al.: Gradient Surgery for Multi-Task Learning, Advances in Neural Information Processing Systems, 33, 5824-5836, 2020.

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.