Revisions of Manuscript: ESSD-2025-192

Title: Spatially adaptive estimation of multi-layer soil temperature at a daily time-step across China during 2010-2020

Author(s): Xuetong Wang, Liang He, Peng Li, Jiageng Ma, Yu Shi, Qi Tian, Gang Zhao, Jianqiang He, Hao Feng, Hao Shi, Qiang Yu

Dear Reviewer,

We sincerely thank you for your thoughtful comments and constructive suggestions on our manuscript. We have carefully revised the manuscript in response to your feedback, with all changes clearly marked using track changes. In the revised manuscript and accompanying supplementary materials, modifications are highlighted in blue for ease of reference.

Below, we provide a detailed, point-by-point response to each of your comments. For clarity, your original remarks are shown in *italics*, followed by our corresponding replies. We have made every effort to address all concerns comprehensively and to improve the scientific rigor, clarity, and overall quality of the manuscript.

We sincerely appreciate the time and effort you invested in reviewing our work, and we believe the revisions have significantly improved the manuscript.

Reviewer Comment 1:

The authors state that the sites were randomly split into training (70%), validation (20%), and test (10%) sets. For geospatial data like soil temperature, which exhibits strong spatial autocorrelation, this random splitting is a critical methodological flaw. It almost certainly leads to "data leakage", where test sites are geographically close to training sites. Consequently, the model can achieve high performance on the test set even though it did not learn the true underlying relationships between predictors and Ts. This means the model's ability to generalize to new, un-sampled areas is not being properly evaluated. The reported performance metrics (e.g., $R^2 > 0.93$ in Fig. 5) are therefore very likely to be significantly inflated and overly optimistic.

The authors should implement a more rigorous validation scheme that accounts for spatial autocorrelation. A spatial block cross-validation approach is strongly recommended.

Response to Reviewer Comment 1:

We sincerely thank the reviewer for this insightful comment. We fully agree that random splitting of sites into training, validation, and test sets may lead to spatial data leakage due to the strong spatial autocorrelation of soil temperature. This could indeed result in overly optimistic performance metrics and an inaccurate assessment of the model's spatial generalization ability.

In response, we have revised our methodology by adopting a spatial block cross-validation scheme to partition the data. Specifically, observation sites were grouped into spatial blocks, and the cross-validation was conducted across these blocks rather than through random splits. This approach ensures that geographically adjacent sites are not simultaneously included in both training and testing subsets, thereby providing a more rigorous and realistic evaluation of model generalization to un-sampled regions.

We have retrained and re-evaluated the XGBoost models using this revised validation strategy. The updated results, along with a detailed description of the method, are now presented in the manuscript.

Revised Text (L318–336):

To rigorously account for the strong spatial autocorrelation of T_s and avoid potential data leakage between training and testing subsets, we employed a spatial block cross-validation scheme rather than random splitting. Specifically, within each rotated quadtree grid, observation sites were grouped into spatial blocks based on their geographic coordinates: station latitude and longitude were each divided by 1° and floored to integer values, and stations sharing the same index were assigned to the same block. This ensured that samples within the same spatial block were not simultaneously assigned to both the training and testing subsets, thereby avoiding data leakage due to spatial autocorrelation and enabling a more reliable evaluation of the model's generalization capability.

Within each spatial grid, the data were partitioned into training (90%) and testing (10%) subsets at the block level. The training subset was further subjected to 10-fold spatial block cross-validation using GridSearchCV to optimize three key hyperparameters: the number of trees (n_estimators), maximum tree depth (max_depth), and learning rate (learning_rate). Detailed parameter settings are provided in Appendix Table S1. The hyperparameter set that yielded the lowest average validation error across the ten folds was selected as optimal. The final model, retrained on the full training set with these parameters, was then evaluated on the held-out testing blocks to assess its generalization ability and examine potential overfitting within each grid.

Reviewer Comment 2:

The manuscript claims that the generated dataset accurately captures the spatial distribution of Ts. However, the evidence provided is the high R^2 (and low RMSE) of the daily time series at individual stations. These temporal variations are heavily dominated by the seasonal cycle, which is easy for any model to capture using predictors like air temperature. A high temporal R^2 does not prove that the model correctly reproduces the spatial gradients across China. I suggest the authors conduct a spatial-only validation, using mean Ts (for the whole year and for specific seasons) across the sites.

Response to Reviewer Comment 2:

We thank the reviewer for this constructive suggestion. We agree that high temporal R^2 at individual stations mainly reflects the ability to capture seasonal variations and may not sufficiently demonstrate the model's capacity to reproduce spatial gradients. Following the reviewer's advice, we conducted a spatial-only validation using annual mean T_s across all stations. The results are presented in Figure 1, which compares the estimated and observed annual mean T_s at depths from 0–40 cm. Each point represents the annual mean T_s at a single site. The results indicate high correlations (R^2 = 0.995–0.997) and low errors (RMSE = 0.26–0.37 K; MAE = 0.19–0.28 K), demonstrating that the generated dataset reliably captures the spatial distribution of T_s across sites. These additional analyses provide strong evidence that our dataset reproduces both temporal dynamics and spatial gradients of T_s across China.

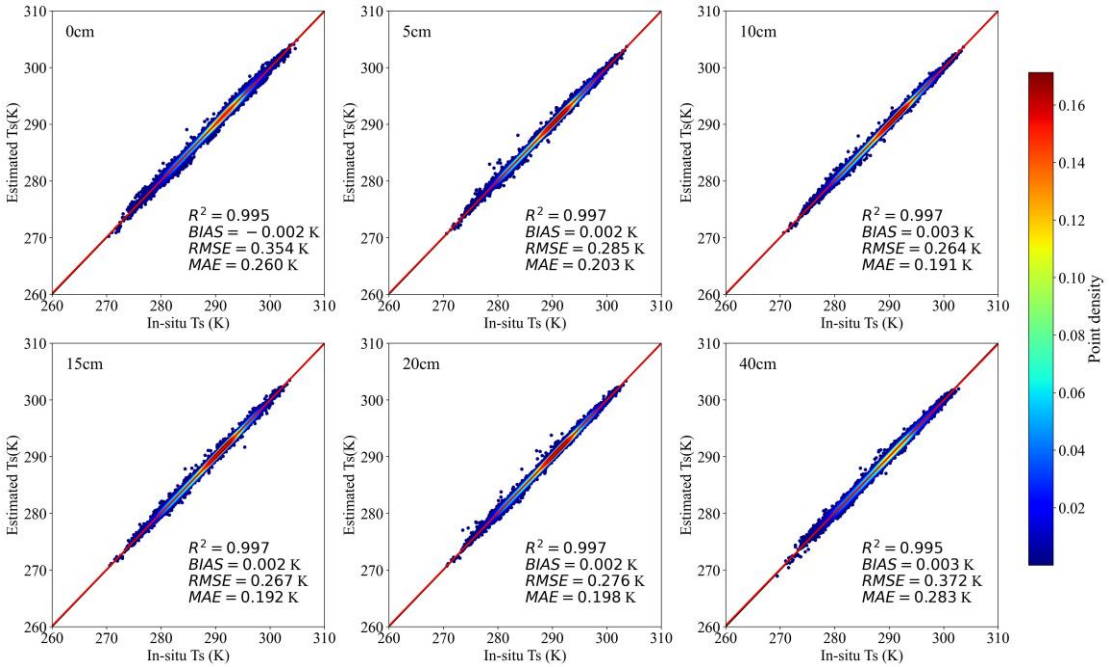

Figure 1. Validation of spatial patterns of annual mean T_s at different soil depths across China.

Reviewer Comment 3:

The results indicate that model performance is worse at the surface (0 cm) and improves at intermediate depths (e.g., 5-20 cm), as shown in Figures 4-7. This is a counter-intuitive result given the layer-cascading methodology, where the prediction for a deeper layer depends on the prediction from the layer above. This structure implies that errors from the surface prediction should propagate downwards, theoretically leading to a degradation of performance with depth. This apparent paradox should be discussed.

Response to Reviewer Comment 3:

As the reviewer correctly noted, our revised modeling results reveal clear depth-dependent variations in prediction accuracy. Overall, acceptable performance was achieved across all depths. Errors were relatively larger at the 0 cm surface layer, whereas predictions at 5 cm and 10 cm depths showed improved accuracy compared to the surface. With further increases in depth (20–40 cm), errors tended to accumulate, and this pattern was particularly evident in summer and winter.

This phenomenon can be explained by the physical characteristics of soil temperature dynamics. The surface layer is strongly influenced by high-frequency environmental disturbances such as radiation, precipitation, and evapotranspiration, which elevate the noise level and complicate accurate prediction. In contrast, intermediate layers benefit from the buffering effects of thermal diffusion and soil heat capacity, which dampen short-term fluctuations and make temperature variations more stable and thus more predictable. At greater depths, however, cascading errors are gradually propagated and

amplified, resulting in reduced accuracy. We have revised the manuscript to include a detailed discussion on the rationale behind this result.

Revised Text (L632-662):

Our results (Figures 8 and 9) show that model accuracy varies across different soil depths, with additional influences from season and land use. Accuracy is relatively lower at the surface (0 cm), improves at intermediate depths (5–10 cm), and then declines again at greater depths (20–40 cm). This depth-dependent pattern can be explained by the physical characteristics of soil temperature. Surface soil temperature is highly sensitive to short-term meteorological fluctuations such as radiation, precipitation, and evapotranspiration, leading to greater spatiotemporal variability and larger prediction errors. In contrast, intermediate soil layers benefit from the buffering effects of thermal diffusion and soil heat capacity, which dampen high-frequency fluctuations and stabilize the relationship between predictors and T_s, thereby improving performance at these depths. At greater depths, however, surface-level errors propagate downward through the cascading framework, resulting in reduced accuracy—particularly during summer and winter.

Seasonal changes and variations in land cover further contribute to differences in estimation accuracy. As shown in Figures 8 and 9, the model exhibits higher accuracy in spring and autumn, whereas its performance tends to decline during summer and winter. During summer, dense vegetation growth and canopy closure reduce the influence of surface–atmosphere energy exchanges on T_s, weakening the correlation between canopy temperature and subsurface T_s (Kropp et al., 2020; Cui et al., 2022). In winter, snow cover introduces a suite of confounding effects: high surface albedo reduces net radiation (Loranty et al., 2014; Li et al., 2018), while snow acts as an insulator, limiting the soil's response to cold air incursions (Zhang, 2005; Myers-Smith et al., 2015). Additionally, low temperatures lead to soil water freezing, which alters the soil's thermal conductivity and heat storage capacity. These factors, together with frequent freeze–thaw cycles, introduce complex nonlinear dynamics in T_s that increase modeling uncertainty (Li et al., 2023a; Imanian et al., 2024). While our multi-source adaptive modeling framework performs well across depths, it does not explicitly account for the physical mechanisms of vertical heat transfer. Future research could explore deep learning models that are capable of learning complex spatiotemporal features and improving the physical interpretability of T_s variations across time, space, and depth.

Reference

Cui, X., Xu, G., He, X., and Luo, D.: Influences of seasonal soil moisture and temperature on vegetation phenology in the Qilian Mountains, Remote Sens., 14, 3645, https://doi.org/10.3390/rs14153645, 2022.

Imanian, H., Mohammadian, A., Farhangmehr, V., Payeur, P., Goodarzi, D., Hiedra Cobo, J., and Shirkhani, H.: A comparative analysis of deep learning models for

soil temperature prediction in cold climates, Theor. Appl. Climatol., 155, 2571–2587, https://doi.org/10.1007/s00704-023-04781-x, 2024.

Kropp, H., Loranty, M. M., Natali, S. M., Kholodov, A. L., Rocha, A. V., Myers-Smith, I., Abbot, B. W., Abermann, J., Blanc-Betes, E., Blok, D., Blume-Werry, G., Boike, J., Breen, A. L., Cahoon, S. M. P., Christiansen, C. T., Douglas, T. A., Epstein, H. E., Frost, G. V., Goeckede, M., Høye, T. T., Mamet, S. D., O'Donnell, J. A., Olefeldt, D., Phoenix, G. K., Salmon, V. G., Sannel, A. B. K., Smith, S. L., Sonnentag, O., Vaughn, L. S., Williams, M., Elberling, B., Gough, L., Hjort, J., Lafleur, P. M., Euskirchen, E. S., Heijmans, M. M., Humphreys, E. R., Iwata, H., Jones, B. M., Jorgenson, M. T., Grünberg, I., Kim, Y., Laundre, J., Mauritz, M., Michelsen, A., Schaepman-Strub, G., Tape, K. D., Ueyama, M., Lee, B.-Y., Langley, K., and Lund, M.: Shallow soils are warmer under trees and tall shrubs across arctic and boreal ecosystems, Environ. Res. Lett., 16, 015001, https://doi.org/10.1088/1748-9326/abc994, 2020.

Li, Q., Ma, M., Wu, X., and Yang, H.: Snow cover and vegetation-induced decrease in global albedo from 2002 to 2016, J. Geophys. Res. Atmospheres, 123, 124–138, https://doi.org/10.1002/2017JD027010, 2018.

Li, X., Zhu, Y., Li, Q., Zhao, H., Zhu, J., and Zhang, C.: Interpretable spatio-temporal modeling for soil temperature prediction, Front. For. Glob. Change, 6, 1295731, https://doi.org/10.3389/ffgc.2023.1295731, 2023.

Loranty, M. M., Berner, L. T., Goetz, S. J., Jin, Y., and Randerson, J. T.: Vegetation controls on northern high latitude snow-albedo feedback: Observations and CMIP 5 model simulations, Glob. Change Biol., 20, 594–606, https://doi.org/10.1111/gcb.12391, 2014.

Myers-Smith, I. H., Elmendorf, S. C., Beck, P. S. A., Wilmking, M., Hallinger, M., Blok, D., Tape, K. D., Rayback, S. A., Macias-Fauria, M., Forbes, B. C., Speed, J. D. M., Boulanger-Lapointe, N., Rixen, C., Lévesque, E., Schmidt, N. M., Baittinger, C., Trant, A. J., Hermanutz, L., Collier, L. S., Dawes, M. A., Lantz, T. C., Weijers, S., Jørgensen, R. H., Buchwal, A., Buras, A., Naito, A. T., Ravolainen, V., Schaepman-Strub, G., Wheeler, J. A., Wipf, S., Guay, K. C., Hik, D. S., and Vellend, M.: Climate sensitivity of shrub growth across the tundra biome, Nat. Clim. Change, 5, 887–891, https://doi.org/10.1038/NCLIMATE2697, 2015.

Zhang, T.: Influence of the seasonal snow cover on the ground thermal regime: An overview, Rev. Geophys., 43, https://doi.org/10.1029/2004RG000157, 2005.

Reviewer Comment 4:

In the VIF analysis (Fig. S1), sand, silt, and clay percentages were included. As these three variables are compositional and should sum to a constant (100%), they are perfectly collinear by definition. This should result in an infinite (or extremely large) VIF values. However, the reported VIFs are relatively low (5.6 to 10). This discrepancy is concerning and suggests a methodological error.

Response to Reviewer Comment 4:

We thank the reviewer for pointing out this important issue. The reviewer is correct that sand, silt, and clay percentages are compositional variables that sum to 100% and are therefore perfectly collinear by definition. When variables are perfectly collinear, VIF cannot be correctly computed, as the underlying regression matrix becomes singular. Including all three variables simultaneously in the VIF analysis was therefore inappropriate, and we acknowledge that this led to misleading values (5.6–10) instead of extremely high or infinite VIFs.

In the revised manuscript, we have addressed this issue by excluding silt from the VIF analysis, since the three variables contain redundant information. This adjustment removes perfect collinearity and allows the VIF analysis to be correctly applied. The updated VIF results are now reported in the Supplementary Material (Fig. S2), and the corresponding text has been revised accordingly.

We further retrained the XGBoost models using the revised set of predictor variables and a spatial block cross-validation data partitioning strategy, and regenerated new data products to ensure the consistency and robustness of the analysis results. We sincerely appreciate the reviewer's suggestion, which has enabled us to improve the methodological rigor and reliability of our study.

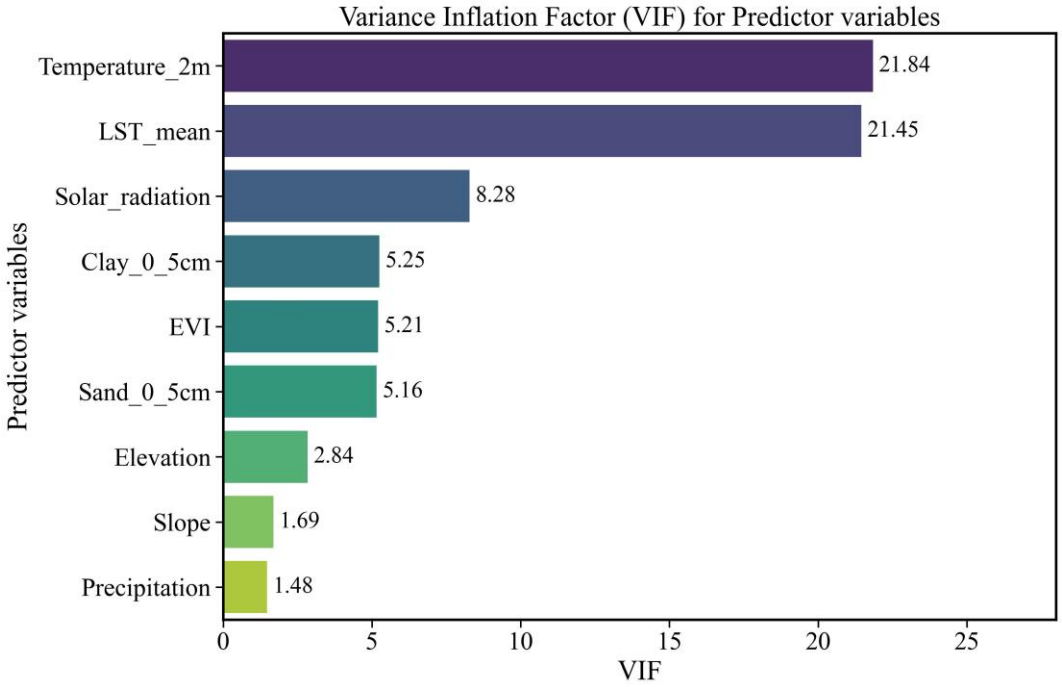

Figure S2. Variance Inflation Factor (VIF) of predictor variables

Revised Text (L244-254):

Multicollinearity among multiple source variables may affect the robustness of the models. Therefore, we rigorously evaluated the multicollinearity among the independent variables using the variance inflation factor (VIF) before modeling to remove highly correlated variables. The VIF is a diagnostic statistic used to quantify the degree of multicollinearity by measuring how much the variance of a regression coefficient is inflated due to correlations with other predictors (Akinwande et al., 2015). It is calculated as:

$$VIF_i = \frac{1}{1 - R_i^2} \quad (1)$$

where R_i^2 is the coefficient of determination obtained by regressing the i-th predictor against all other predictors. Variables with VIF exceeding 10 are generally considered severely multicollinear and should be removed.

Reference

Akinwande, M. O., Dikko, H. G., and Samson, A.: Variance inflation factor: As a condition for the inclusion of suppressor variable(s) in regression analysis, Open J. Stat., 5, 754–767, https://doi.org/10.4236/ojs.2015.57075, 2015.

Reviewer Comment 5:

The model uses solar radiation as a predictor but omits downward longwave radiation (LWD). Considering that LWD is a critical driver of the surface energy balance

(particularly for nighttime and winter temperatures) and that LWD has been identified as a main driver of Ts trends in process-based models (Peng et al., 2016, https://doi.org/10.5194/tc-10-179-2016), I suggest the authors include LWD as a predictor, or provide a strong justification for its exclusion.

Response to Reviewer Comment 5:

We thank the reviewer for this valuable suggestion. Following the reviewer's comment, we incorporated downward longwave radiation (LWD) from ERA5 as a candidate predictor and evaluated its multicollinearity with other variables. The analysis revealed that LWD is highly collinear with solar radiation (revised Fig. S1). Considering that our study focused on daily mean T_s, the additional contribution of LWD was limited at the daily scale, as its effect on the surface energy balance was already largely captured by solar radiation. For these reasons, we excluded LWD from the final modeling to avoid redundancy and potential instability in the regression framework. Importantly, the inclusion or exclusion of LWD did not materially change the results or conclusions of our study.

This clarification has been added to the revised manuscript, and the updated figure illustrating the collinearity analysis is provided in the Supplementary Material.

Revised Text (L158-164):

In addition, both net solar radiation and downward longwave radiation (LWD) were considered. Net solar radiation directly represents the shortwave energy absorbed by the land surface and serves as the primary driver of the daytime surface energy budget, whereas LWD plays a particularly important role under nighttime and winter conditions by regulating surface heat loss through the longwave radiation balance. Together, they jointly control the surface energy balance and directly drive the spatiotemporal dynamics of T_s (Peng et al., 2016).

Revised Text (L255-268):

Based on the VIF analysis, we applied the following adjustments to the predictor set. Accordingly, some variables were excluded due to severe multicollinearity or redundancy. Specifically, sand, silt, and clay are compositional variables whose proportions sum to 100%, leading to perfect collinearity. To reduce redundancy, we removed silt while retaining sand and clay. In addition, LWD was found to be highly correlated with net solar radiation at the daily mean scale (Fig. S1) and was therefore excluded from the final modeling.

In contrast, although the daily mean LST (LST_mean) and air temperature also exhibited strong collinearity, with VIF values exceeding 10 (Fig. S2), we decided to retain both. This decision reflects their physical distinctness and complementary information: LST_mean provides higher spatial resolution (1 km), whereas air temperature offers broader meteorological consistency (9 km). Such differences are particularly important in complex ecosystems such as forests, where canopy structure

and biological processes substantially influence thermal dynamics (Liu et al., 2025).

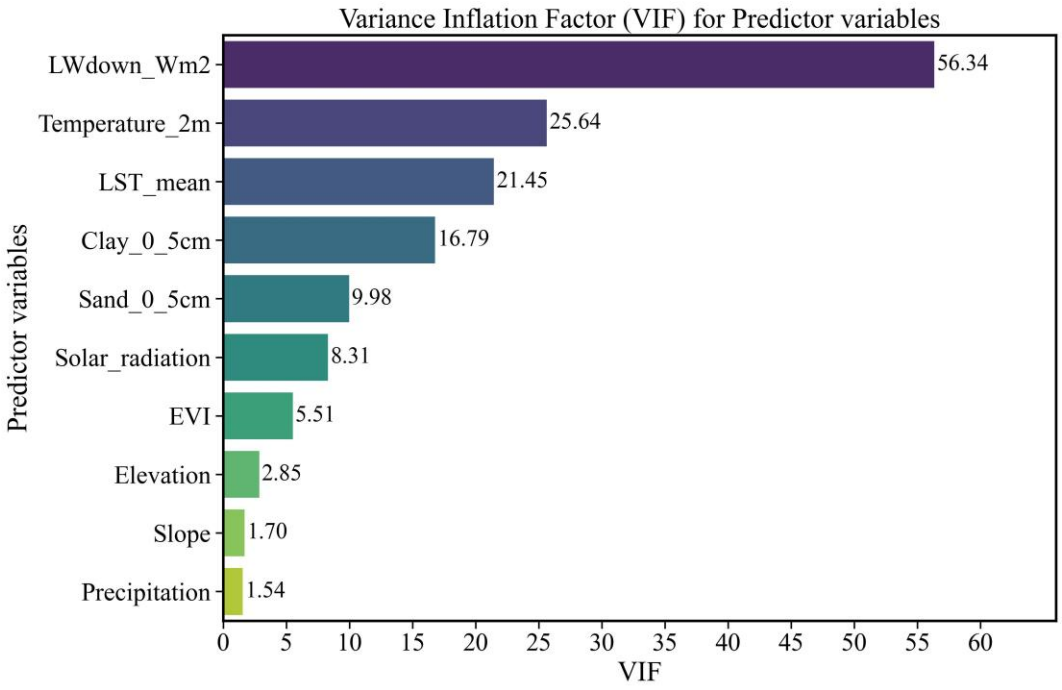

Figure S1. Variance Inflation Factor (VIF) of predictor variables (with LWD)

Reference

Liu X., Li Z.-L., Duan S.-B., Leng P., and Si M.: Retrieval of global surface soil and vegetation temperatures based on multisource data fusion, Remote Sens. Environ., 318, 114564, https://doi.org/10.1016/j.rse.2024.114564, 2025.

Peng, S., Ciais, P., Krinner, G., Wang, T., Gouttevin, I., McGuire, A. D., Lawrence, D., Burke, E., Chen, X., Decharme, B., and others: Simulated high-latitude soil thermal dynamics during the past 4 decades, The Cryosphere, 10, 179–192, 2016.

Reviewer Comment 6:

Line 490 "Notably, RMSE at the surface (0 cm) is slightly lower than at 40 cm, possibly due to stronger direct influences from surface cover and meteorological conditions." – This is not the case for Fig. 12 cd. Furthermore, making this statement based on only a few sites is not adequate.

Response to Reviewer Comment 6:

We thank the reviewer for the valuable observation. We agree that the RMSE at 0 cm is not consistently lower than at 40 cm across all stations. Our original statement was overly generalized based on a limited number of sites and may have caused confusion. We have revised the text accordingly to avoid overinterpretation.

Revised Text (L505-506):

Site-level accuracy was evaluated using RMSE, which ranged from 0.84 K to 1.80 K

across both depths, indicating strong agreement between predicted and observed values.

Reviewer Comment 7:

Line 535 "Figure. S5 demonstrates that LST is more effective than air temperature in detecting spatial variations in surface Ts in sparsely vegetated areas" – I do not see how this conclusion can be derived from Fig. S5.

Response to Reviewer Comment 7:

We thank the reviewer for this valuable comment. We agree that Fig. S5 alone does not provide direct evidence that LST is more advantageous than air temperature in sparsely vegetated areas. In response, we have revised the text in the manuscript, removed the description related to Fig. S5, and added supporting evidence from Figs. S7 and S8 to more robustly substantiate this conclusion.

Revised Text (L556-560):

As shown in Figs. S7 and S8, incorporating LST as an input variable, relative to using only air temperature, significantly enhances overall modeling accuracy and improves performance across sites with different land cover types, with the most pronounced improvements observed in barren land areas.

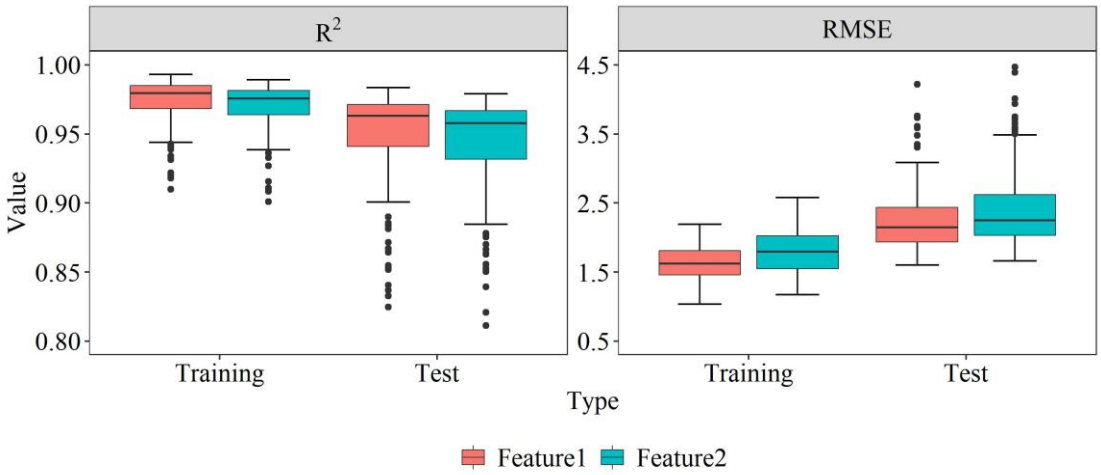

Figure S7. Comparison of Modeling Accuracy with Different Feature Variables (Feature1 represents using both air temperature and LST together with other feature variables, while Feature 2 represents using only air temperature together with other feature variables)

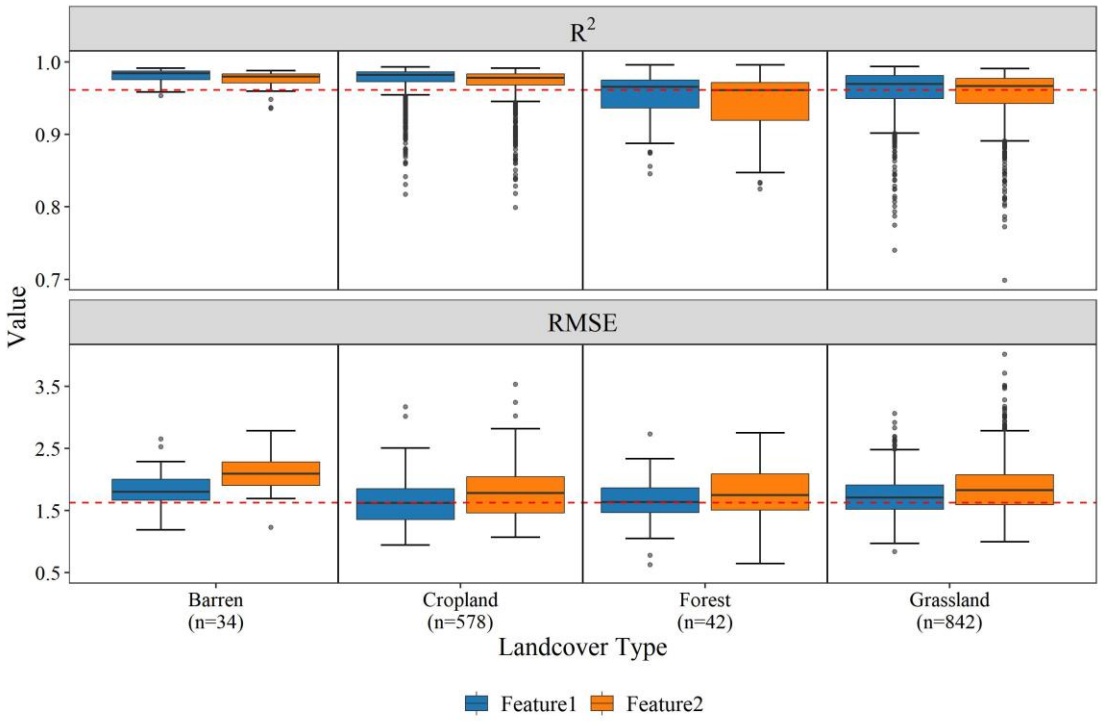

Figure S8. Differences in model accuracy across land cover types under different feature variable combinations. (Feature1 represents using both air temperature and LST together with other feature variables, while Feature 2 represents using only air temperature together with other feature variables)