Response to Referee Comments on Manuscript ESSD-2026-447

Dear Editor and Referees,

We sincerely thank the editor and the two anonymous referees for their careful evaluation of our manuscript and for the constructive comments and suggestions. We have carefully addressed all comments and prepared a substantially revised version of the manuscript. The major revisions include: (1) a comprehensive re-examination and targeted quality control of the ALT observational database, followed by a complete rerun of model training, validation, uncertainty analysis, and dataset production; (2) the addition of 10° × 10° spatial block cross-validation and temporally blocked validation using newly compiled observations from 2021–2024; (3) an area-of-applicability analysis to identify regions involving environmental extrapolation; (4) revised terminology and interpretation of the site-resampling uncertainty analysis; (5) additional analyses including SoilGrids depth sensitivity and SHAP-based predictor importance; and (6) improved methodological documentation and a more rigorously harmonized comparison with existing hemispheric ALT products.

These revisions have substantially strengthened the evaluation, interpretation, and reproducibility of the dataset. Our point-by-point responses are provided below.

Referee #1:

Comments 1:

All records come from GTN-P, of which CALM is the active layer sub-network. Although CALM is currently the only authoritative public data source for active layer thickness in permafrost regions, the published values do not strictly distinguish between active layer thickness and the depth to the permafrost table, two fundamentally different concepts. In continuous permafrost areas the two coincide numerically, but in discontinuous permafrost areas the active layer thickness is determined by the

maximum seasonal frost depth, while the permafrost table lies deeper than the base of the active layer because of talik layers in between. As a rule of thumb, for sites with reported values between 3 and 5 m, part of these values may in fact represent the depth to the permafrost table, and values larger than 5 m should almost certainly be interpreted as the permafrost table depth rather than the active layer thickness. I recommend that the authors screen and select the sites that genuinely represent active layer thickness for model training, using the auxiliary documentation available from the public data sources. For instance, in the supplementary material provided by the authors, some sites report active layer thicknesses close to 10 m, which is theoretically impossible.

Response 1:

Thank you for raising this important point. We agree that exceptionally large values reported in CALM/GTN-P require careful interpretation, particularly because maximum thaw depth and classical active-layer thickness may diverge under strongly degraded or talik-affected conditions. We therefore re-examined the observational database and revised the quality-control procedure.

We did not apply a simple fixed upper threshold to remove all deep observations, because CALM/GTN-P reports these measurements within its active-layer/annual-thaw-depth framework and some high values may represent genuine deep-thaw conditions. Applying an arbitrary cutoff would also truncate the upper tail of the calibration data and could bias the model toward shallower ALT conditions. Instead, exceptionally deep records were subjected to targeted quality control. Isolated extreme values, abrupt discontinuities relative to adjacent years, and extreme records from very short site series that lacked sufficient temporal support were removed. We then repeated the complete model training, validation, and mapping workflow using the revised observational dataset.

We have also clarified in Section 2.2 that ALT is used operationally following the CALM/GTN-P convention and refers to annual maximum thaw depth. In locations where the ground fully refreezes each winter, this corresponds to classical active-layer thickness, whereas in strongly degraded or talik-affected permafrost some deep values

may instead approximate the depth to the permafrost table. The corresponding limitation is also discussed in Section 3.6.

Comments 2:

For the evaluation of the temporal simulation performance, why were only 68 sites used rather than all sites with a monitoring record longer than 10 years (according to the training sample data provided by the authors, there are 101 such sites)? I cannot tell whether this is an oversight as described in the text, or whether the other sites performed poorly in the temporal validation and were deliberately removed. If it is the latter, I would urge the authors to report the statistics honestly, so as not to mislead readers. This is a data paper; ALT observations are inherently scarce, and a somewhat lower accuracy of the derived product would not in itself affect the publishability of the work. Honest reporting of the accuracy assessment is what matters most.

Response 2:

Thank you for pointing this out. We apologize that the site-selection criteria for the long-term temporal evaluation were incompletely described in the original manuscript. In the original analysis, sites were required to satisfy two criteria: (1) at least 10 valid annual ALT observations and (2) an observation span exceeding 15 years during 2000–2020. However, only the first criterion was stated in the manuscript, which understandably created the impression that all sites with more than 10 observations should have been included.

The additional observation-span criterion was applied from the outset to ensure that site-level temporal trends were estimated from sufficiently long records. No sites were excluded on the basis of model performance or temporal-validation results. We have now corrected the Methods to explicitly state both criteria and clarified that all sites satisfying these criteria were included in the temporal evaluation.

Following the updated quality-control procedure and reconstruction used in the revised manuscript, 60 sites now satisfy both criteria and are included in the long-term temporal assessment.

Comments 3:

In the spatial validation of the training samples the authors adopted LOSO-CV, but environmentally similar neighbouring sites usually remain in the training set, so LOSO largely measures interpolation ability and its results are inevitably optimistic. I suggest that, in addition to LOSO-CV, the authors perform spatial block cross-validation or region-level cross-validation, map the area of applicability (AOA) of the final model following established practice (e.g. Meyer & Pebesma, 2022, https://doi.org/10.1111/2041-210X.13650), and either mask out or clearly flag the parts of the released product where the predictions fall outside the feature space covered by the calibration data, i.e. where they are extrapolations.

Response 3:

Thank you for this important suggestion. We have added a 10° × 10° spatial block cross-validation and used it as the primary framework for comparing candidate models and selecting the final LightGBM–XGBoost ensemble. LOSO-CV was retained as a complementary site-level validation. As expected, performance under spatial block cross-validation was lower than under LOSO-CV, confirming that the former provides a more stringent assessment of spatial generalization.

We also added an area-of-applicability (AOA) analysis following the general framework of Meyer and Pebesma (2021), with details provided in the revised Methods. Because several predictors vary annually, AOA was evaluated separately for 2000–2024. Most of the mapped permafrost region remained within the AOA, whereas localized areas were identified as environmental extrapolation.

Rather than masking these pixels from the released ALT product, we chose to retain the complete predictions and clearly flag extrapolated areas. Annual binary AOA layers are now provided together with the ALT dataset, where AOA = 1 indicates predictions within the environmental domain represented by the calibration data and AOA = 0 indicates environmental extrapolation. An AOA-frequency layer summarizing the proportion of years within the AOA during 2000–2024 is also provided. The corresponding Methods, Results, limitations, and dataset documentation have been

revised accordingly.

Comments 4:

SoilGrids provides soil properties for several standard depth intervals (0-5, 5-15, 15-30, 30-60, 60-100 and 100-200 cm), but the manuscript never states which interval was used for bulk density, coarse fragments and silt. ALT spans from tens of centimetres to several metres across the sites. Please state the depth intervals used, justify the choice, and ideally test the sensitivity of the mapped product to this choice.

Response 4:

Thank you for pointing this out. We have now clarified the SoilGrids depth treatment in the revised manuscript. SoilGrids provides bulk density, coarse fragment content, sand, silt, and clay for six standard depth intervals (0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm). For each soil variable, we calculated a thickness-weighted mean over the full 0–200 cm profile and used this value as the static soil predictor. We selected 0–200 cm because it represents the deepest consistently available SoilGrids profile and provides an integrated characterization of near-surface substrate conditions. These soil variables are intended to represent broad background properties relevant to active-layer thermal and hydrological conditions rather than depth-resolved properties throughout the entire seasonally thawed layer, which can locally exceed 2 m.

We also conducted a sensitivity analysis using thickness-weighted soil properties integrated over 0–30, 0–100, and 0–200 cm while keeping the final LightGBM–XGBoost ensemble and spatial block cross-validation framework unchanged. Model performance was nearly identical for the 0–100 and 0–200 cm configurations (R^2 = 0.722 vs. 0.720; RMSE = 39.36 vs. 39.47 cm; CCC = 0.853 vs. 0.849), whereas the 0–30 cm configuration showed slightly lower performance (R^2 = 0.696; RMSE = 41.19 cm; CCC = 0.838). These results indicate that model performance is only weakly sensitive to whether soil properties are integrated over 0–100 or 0–200 cm. We therefore retained the 0–200 cm profile because it provides the most complete available representation of the soil column while yielding predictive performance comparable to the 0–100 cm configuration. The depth treatment and rationale have been clarified in

Section 2.3.2, and the sensitivity results have been added to Table S2.

Comments 5:

Among all input variables of this study, Biome appears to be the only categorical one. Some of the models used by the authors (e.g. Random Forest and GBDT) do not support categorical variables and require special treatment. It is not clear to me how Biome was handled in each model; for global-scale spatial prediction the effect is probably considerable, and I suggest the authors provide an explanation in the main text.

Response 5:

Thank you for pointing this out. In this study, Biome was treated as a categorical predictor and transformed using one-hot encoding prior to model training. Each biome class was therefore represented by a separate binary indicator, and the same encoded representation was used consistently across all five candidate models, including RF, GBDT, XGBoost, CatBoost, and LightGBM. This treatment avoids imposing an artificial ordinal relationship among biome classes and ensures that model comparisons are based on a consistent feature representation. We have now clarified this procedure in the main text in Section 2.4.2.

Comments 6:

The variable screening in Section 3.1 is based on correlation analysis performed before modelling, and no importance measure (such as SHAP values) is reported anywhere in the manuscript, so I cannot judge which predictors actually drive the predictions. This also bears on the resolution of the final product. Table S1 lists the native resolution of each predictor, and several of the strongest ones are much coarser than 1 km (e.g. ERA5-Land (SR, SDE) at 0.1 degrees, and WorldClim (MAT, MAP, TI) at 2.5 arc-minutes). Three of the five predictors with the highest correlation with ALT belong to this group. Resampling these layers onto a 1 km grid does not create 1 km information. I therefore suggest that the effective resolution of the product be inferred from the native resolution of the most important predictors identified by a SHAP analysis; if the dominant predictors are coarser than 1 km, running the model on a 1 km

grid does not mean the product truly has 1 km resolution.

Response 6:

Thank you for this important comment. We have added a SHAP analysis for the final LightGBM–XGBoost ensemble to quantify the contribution of individual predictors to model predictions. The results show that elevation was the most influential predictor, accounting for 34.2% of the total mean absolute SHAP contribution, followed by bulk density (13.1%), MAT (12.6%), TI (8.8%), and slope (7.6%) (Fig. S7). Importantly, elevation and slope were derived from the 15 arc-second GEBCO dataset, providing sub-kilometre-scale topographic information, while soil predictors also contributed substantially to the model. At the same time, several climatic predictors, including MAT and TI, originate from coarser-resolution datasets.

We have therefore revised the manuscript in Section 3.6 to distinguish the 1 km output grid spacing from the effective spatial information content of the product. We now explicitly state that resampling coarser predictors to 1 km does not create additional fine-scale information and that the effective spatial detail of the reconstructed ALT reflects the combined native resolutions and spatial structures of the contributing predictors. Accordingly, the product should be interpreted as a 1 km gridded ALT dataset rather than as evidence that all underlying environmental information is resolved independently at 1 km. To further avoid implying a uniform effective spatial resolution of 1 km, we have also revised the manuscript title from "A 1 km resolution dataset of Northern Hemisphere permafrost active layer thickness (2000–2024)" to "**A 1 km gridded dataset of Northern Hemisphere permafrost active layer thickness (2000–2024)**." We have also clarified in Section 2.3 that Pearson correlation analysis was used only for preliminary predictor screening and redundancy assessment.

Comments 7:

The manuscript does not clearly explain how the hyperparameters of each model were determined. Please add this information.

Response 7:

Thank you for pointing this out. We have now clarified the determination and use

of model hyperparameters in the revised Methods. The model configurations were established during preliminary model development and were fixed before formal model evaluation. The same parameter settings were subsequently retained across the spatial block cross-validation, LOSO-CV, and temporal blocked validation experiments to ensure consistent model comparison. The main hyperparameter settings adopted for RF, GBDT, XGBoost, CatBoost, and LightGBM are now provided in Table S3 for reproducibility.

Comments 8:

The data product covers 2000-2024, but the calibration data end in 2020. In my view, extending the product beyond the calibration period would be acceptable if later spatial block or region-level cross-validation shows high accuracy over most of the study area; if the validation accuracy is poor, I see no need to deliberately extend the coverage, because the accuracy of the extended years is likely to be even lower, unless the authors can provide a sound justification.

Response 8:

Thank you for raising this important concern. In the revised manuscript, we incorporated newly available ALT monitoring records from 2021–2024 and used them first as an independent temporal validation dataset. Specifically, the ensemble model was trained exclusively using observations from 2000–2020, while all available observations from 2021–2024 were withheld from model fitting and used only for evaluation. Across all withheld 2021–2024 records, including both previously observed and newly observed sites, the model achieved an R^2 of 0.693, an RMSE of 26.36 cm, an MAE of 17.55 cm, a CCC of 0.802, and a mean bias of 0.51 cm. When the evaluation was restricted to sites already represented during 2000–2020, performance increased to an R^2 of 0.826, an RMSE of 19.74 cm, an MAE of 14.06 cm, a CCC of 0.895, and a mean bias of 0.13 cm. These results indicate that the relationships learned from the 2000–2020 observations remained transferable to the subsequent period, although prediction errors were larger when temporal transfer was combined with spatial extrapolation to newly observed sites.

After completing this independent temporal validation, the final LightGBM–XGBoost ensemble was refitted using the complete quality-controlled ALT observation dataset for 2000–2024, including the newly added 2021–2024 records, and this refitted model was used to generate the released annual ALT maps. Therefore, the 2021–2024 layers in the revised product are no longer predictions generated purely beyond the calibration period. We have revised the Methods, Results, and limitations accordingly.

Comments 9:

CCC is used as an accuracy metric but never defined. Please define it at first occurrence.

Response 9:

Thank you for noting this omission. We have now defined Lin's concordance correlation coefficient (CCC) at its first occurrence in Section 2.4.2 and added its equation and interpretation. CCC jointly evaluates correlation and agreement with the 1:1 line, with values approaching 1 indicating stronger concordance between observed and predicted ALT.

Referee #2:

Comments 1:

I am not convinced that the current temporal validation adequately supports the extension of the dataset to 2021–2024. In the LOSO framework, all observations from one site are excluded, but observations from the same years at other sites remain in the training dataset. Therefore, this validation primarily evaluates whether temporal variations learned from other locations can be transferred to an unseen site, rather than whether a model trained on historical observations can reliably predict future years. This distinction is important because most of the observations used for model development are from 2000–2020, whereas the final product extends to 2024, and the authors themselves acknowledge that the 2021–2024 maps represent temporal extrapolations. I recommend that the authors perform an additional temporally blocked

validation, for example, by training the model using observations up to a given year and evaluating it against observations from subsequent years. Such an experiment would provide a more direct assessment of the model's temporal extrapolation capability. Without such a test, the reliability of the 2021–2024 product remains difficult to assess.

Response 1:

Thank you for this important comment. In the revised manuscript, we incorporated newly available ALT monitoring records from 2021–2024 and added an explicit temporally blocked validation. The LightGBM–XGBoost ensemble was trained exclusively using observations from 2000–2020, while all available observations from 2021–2024 were withheld from model fitting and used only for evaluation. Across all withheld 2021–2024 records, including both previously observed and newly observed sites, the model achieved an R^2 of 0.693, an RMSE of 26.36 cm, an MAE of 17.55 cm, a CCC of 0.802, and a mean bias of 0.51 cm. When the evaluation was restricted to sites already represented during 2000–2020, performance increased to an R^2 of 0.826, an RMSE of 19.74 cm, an MAE of 14.06 cm, a CCC of 0.895, and a mean bias of 0.13 cm.

This additional experiment provides a direct assessment of temporal transfer from the historical period to 2021–2024. The lower performance for the complete test set reflects the additional challenge of predicting newly observed sites, where temporal extrapolation is combined with spatial generalization.

After completing this independent temporal validation, the final ensemble was refitted using the complete quality-controlled observation dataset for 2000–2024, including the newly added 2021–2024 records, and this refitted model was used to generate the released annual ALT maps. Therefore, the revised 2021–2024 product layers are no longer based purely on extrapolation beyond the calibration period. The Methods, Results, and limitations have been revised accordingly.

Comments 2:

The model appears to substantially underestimate interannual variability. The

reported median fluctuation amplitude ratio (FAR) is only 0.628, indicating that the detrended variability of the reconstructed ALT time series is considerably smaller than that of the observations at many validation sites. This is an important limitation for a dataset provided at an annual temporal resolution. In its present form, the manuscript sometimes gives the impression that the dataset can reliably reproduce interannual ALT variability, whereas the validation results indicate substantial smoothing of year-to-year fluctuations. I recommend that the authors revise the relevant statements throughout the manuscript and make it clear that the product appears to be more suitable for characterizing broad spatial patterns and long-term temporal changes than for investigating anomalous individual years or short-term ALT variability.

Response 2:

Thank you for this important comment. Following your suggestion, we have revised the relevant statements throughout the manuscript to distinguish between long-term temporal behavior and short-term year-to-year fluctuations.

In the revised analysis, LOSO-CV predictions were evaluated across 60 long-term monitoring sites. The mean ALT time series across these sites showed close agreement with the observations (r=0.954, R^2=0.838, RMSE = 1.861 cm), and the observed and predicted long-term Sen's slopes were similar (0.572 and 0.589 cm yr-1, respectively). However, the median FAR was 0.641, confirming that interannual variability at individual sites was substantially smoothed.

We have therefore revised the Results, Abstract, and the "Potential applications and limitations" section to clarify that the dataset is more suitable for characterizing broad spatial patterns and long-term ALT changes than for analyzing anomalous individual years or short-term site-specific variability. Representative site-level time series are also provided in the Supplementary Material to illustrate this limitation directly.

Comments 3:

The uncertainty analysis does not appear to represent the full predictive uncertainty of the ALT product. The authors randomly selected 80% of the monitoring

sites 100 times, refitted the models, and then used the 2.5th and 97.5th percentiles of the resulting predictions as an empirical 95% prediction interval. As implemented, this procedure primarily quantifies the sensitivity of the predictions to changes in the composition of the training-site network. It does not explicitly account for several other potentially important sources of uncertainty, including measurement uncertainty in the ALT observations, residual model error, uncertainty in the predictor datasets, unresolved sub-grid heterogeneity, and uncertainty associated with temporal extrapolation. Therefore, referring to this range as a "95% prediction interval" may give readers the impression that it represents the overall predictive uncertainty of the final product, which is not demonstrated by the current analysis. I recommend that the authors more clearly define the component of uncertainty quantified by this procedure and, if possible, evaluate the empirical coverage of the reported interval using withheld observations. If such calibration is not feasible, the terminology should be revised to more accurately reflect the uncertainty component being represented.

Response 3:

Thank you for this important clarification. We have revised the terminology throughout the manuscript. The previous "95% prediction interval" is now referred to as an "empirical 95% site-resampling interval," and "relative uncertainty" has been revised to "relative site-resampling uncertainty." We have also explicitly clarified that this metric does not account for measurement error, predictor uncertainty, residual model error, unresolved sub-grid heterogeneity, or temporal extrapolation uncertainty.

Comments 4:

The overall LOSO-CV statistics may conceal substantial spatial differences in model performance. The authors report an ensemble of 0.761 and an RMSE of 60.48 cm by pooling all annual observations from the withheld sites. However, sites with longer observation records contribute more samples to these pooled statistics, and substantial environmental differences among Arctic, sub-Arctic, and high-elevation permafrost regions may be obscured by a single hemispheric-scale performance metric. The manuscript also shows that large ALT values tend to be underestimated, suggesting

that model performance is not uniform across the observed ALT range. I recommend that the authors provide additional validation statistics stratified by major geographic region, permafrost zone, or ALT range. Site-level or site-weighted performance metrics could also be reported in addition to the pooled record-level statistics. For a hemispheric-scale data product, users need to understand not only its overall accuracy but also the geographic and environmental conditions under which prediction errors are relatively large.

Response 4:

Thank you for this important comment. In the revised manuscript, we have added site-level performance summaries in addition to the pooled LOSO-CV statistics. For the final LightGBM–XGBoost ensemble, the pooled LOSO-CV yielded an R^2 of 0.780, an RMSE of 34.981 cm, an MAE of 22.053 cm, a CCC of 0.879, and a mean bias of 1.086 cm. We additionally calculated RMSE, MAE, and bias separately for each monitoring site and summarized these metrics using their mean and median values, thereby giving each site equal weight. The median site-level RMSE and MAE were 17.173 and 15.827 cm, respectively, whereas the corresponding mean values were 31.295 and 29.847 cm, indicating that errors were relatively small at many sites but substantially larger at a subset of locations.

We also evaluated performance across the observed ALT range under the more stringent spatial block cross-validation. For ALT <100 cm, the ensemble achieved an RMSE of 26.46 cm, an MAE of 17.19 cm, and a bias of +9.43 cm. Errors increased for ALT of 100–300 cm (RMSE = 62.64 cm; MAE = 52.75 cm; bias = −20.56 cm) and were largest for ALT >300 cm (RMSE = 84.04 cm; MAE = 70.59 cm; bias = −68.12 cm). These results confirm that prediction performance is not uniform across the observed ALT range and that deep-thaw conditions are systematically underestimated.

The revised manuscript now reports both pooled and site-level LOSO-CV statistics, together with ALT-range-stratified spatial block validation results, so that users can better identify the conditions under which prediction errors are larger. These additional results are provided in Tables S4–S5 and Fig. S2–S3.

Comments 5:

The relationship between the reported spatial uncertainty pattern and the distribution of the observations requires further examination. The authors note that monitoring sites are relatively sparse in regions such as interior Siberia and the Tibetan Plateau. However, the uncertainty maps indicate relatively low uncertainty across extensive areas of northern Siberia, whereas high uncertainty is concentrated in more scattered regions. This pattern is not necessarily incorrect, but low disagreement among repeatedly resampled models does not by itself demonstrate high predictive reliability in data-sparse regions. Different model realizations may produce similar predictions even when all of them are extrapolating beyond well-sampled portions of the predictor space. I therefore suggest that the authors examine the relationship between the uncertainty estimates and observation density or distance to the nearest training sites. An analysis of environmental novelty, extrapolation distance, or area of applicability would also be useful for distinguishing regions dominated by interpolation from those requiring substantial extrapolation and would substantially improve the interpretation of the uncertainty product.

Response 5:

Thank you for this important comment. We have added an area-of-applicability (AOA) analysis following the general framework of Meyer and Pebesma (2021). The AOA explicitly evaluates whether the environmental conditions at each prediction pixel are represented within the predictor space covered by the calibration data. Because several predictors vary annually, AOA was calculated separately for each year from 2000 to 2024, and an AOA-frequency layer was additionally generated to summarize the proportion of years in which each pixel remained within the applicability domain.

The AOA analysis showed that most of the mapped permafrost region remained within the environmental domain represented by the calibration data, whereas localized areas were more frequently classified as environmental extrapolation. Importantly, we now interpret the site-resampling uncertainty and AOA as complementary diagnostics: low site-resampling uncertainty indicates that predictions are relatively insensitive to

changes in training-site composition, but does not necessarily imply that the prediction lies within a well-represented portion of environmental space.

We have revised the Methods, Results and limitations accordingly and now provide annual binary AOA layers together with the released ALT product, as well as the 2000–2024 AOA-frequency layer, so that users can explicitly identify areas where predictions involve environmental extrapolation. This distinction also explains why some data-sparse regions can exhibit relatively low site-resampling uncertainty while still requiring separate evaluation of their environmental applicability.

Comments 6:

The comparison with the Westermann et al. and Peng et al. products is useful, but the harmonization procedure should be described more explicitly. The three products show substantial differences in both magnitude and spatial pattern, with pixel-wise RMSE values exceeding 90 cm in both comparisons. The manuscript states that the pixel-wise comparison was conducted using common valid pixels for the 2000–2020 period, but it is less clear whether an equivalent spatial and temporal harmonization was applied when calculating the hemispheric and regional mean ALT values shown in the broader product comparison. Differences in permafrost masks, valid spatial domains, temporal averaging periods, or the treatment of deep thaw and talik conditions could themselves contribute to the reported discrepancies. I agree with the authors that inter-product comparisons should not be interpreted as independent validation. Nevertheless, the procedures used to harmonize the three products should be described in sufficient detail to allow readers to distinguish genuine differences among ALT estimates from differences caused by spatial or temporal domain definitions.

Response 6:

Thank you for this important comment. We have revised the comparison procedure. All three products were first restricted to the common period 2000–2020 and spatially aligned. Hemispheric and regional mean ALT values are now calculated only over pixels that are valid in all three products, so that each product is compared over an identical spatial domain. Regional means are calculated after further intersecting this

common-valid mask with the corresponding regional boundaries. The broader spatial comparison figure has been updated accordingly.

For the pixel-wise comparisons, we retain the pairwise common-valid-pixel approach because these analyses directly compare this study with each reference product separately. Pearson's correlation coefficient, RMSE, and mean difference are therefore calculated using common valid pixels for each product pair.

We have also clarified that remaining discrepancies may reflect not only differences in model structure and input data, but also differences in the definition and treatment of deep thaw and talik conditions. The inter-product comparison is therefore interpreted as an assessment of agreement among products rather than independent validation.

Comments 7:

Several methodological details require further clarification to ensure reproducibility. First, the manuscript should specify which SoilGrids depth interval, or which depth-weighting procedure, was used to derive bulk density, coarse fragment content, sand, silt, and clay. Second, the authors should explain how the categorical biome variable was encoded or handled in each machine-learning algorithm. Third, MATmax and MATmin are defined as the maximum and minimum monthly mean temperatures within a year rather than the absolute annual maximum and minimum air temperatures; therefore, the current terminology may be ambiguous and should be revised or more explicitly defined. Finally, because winter snow depth is calculated from October of the previous year through April of the current year, the authors should clarify how the winter snow-depth predictor for the year 2000 was derived.

Response 7:

Thank you for pointing out these methodological details. We have revised the Methods to clarify all four issues and improve reproducibility.

First, bulk density, coarse fragment content, sand, silt, and clay were derived from the six standard SoilGrids depth intervals (0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm). For each variable, a thickness-weighted mean over 0–200 cm was calculated

and used as the static soil predictor.

Second, Biome was treated as a categorical predictor and transformed using one-hot encoding before model training. The same encoded representation was used consistently across RF, GBDT, XGBoost, CatBoost, and LightGBM, thereby avoiding any artificial ordinal relationship among biome classes.

Third, we agree that the original terminology for MATmax and MATmin could be ambiguous. These variables are now explicitly defined as the annual maximum monthly mean temperature and annual minimum monthly mean temperature, respectively, rather than as absolute annual temperature extremes.

Finally, winter snow depth for a given year was calculated from October of the preceding year through April of the current year. Accordingly, the snow-depth predictor for 2000 was derived from ERA5-Land snow-depth data for October–December 1999 and January–April 2000.

These clarifications have now been incorporated into the revised Environmental variables and Model development sections.