# Author Response to Reviewer 2 Comments

**Authors:**

We sincerely thank you and the reviewers for the time and effort invested in evaluating our manuscript. We greatly appreciate the constructive comments and insightful suggestions, which have contributed to improving the clarity, accuracy, and overall quality of our work. Below, we provide detailed, point-by-point responses to all major and minor comments.

## Reviewer:

This manuscript presents a comprehensive and timely study on the estimation of soil bulk density (BD) using machine learning (ML) and remote sensing data, with a specific focus on temporal changes in Thailand between 2004 and 2009. The authors are to be commended for the extensive comparison undertaken, evaluating six different ML models against a very large benchmark of 76 published pedotransfer functions (PTFs). The use of Bayesian Optimization for hyperparameter tuning represents a rigorous and state-of-the-art approach. The paper is well-structured, the research question is significant, and the results, if validated, would be a valuable contribution to the fields of soil science, remote sensing, and land management.
However, there are several major concerns, primarily methodological, that must be addressed before the manuscript can be considered for publication. The most critical issue is a fundamental contradiction between the model development/validation and its application for temporal analysis, which currently undermines the paper's main conclusions regarding temporal trends.

**Author Response:** We acknowledge the concern regarding a potential contradiction between model development/validation and the temporal analysis, and we address this point in detail under the Major Comments below.

## Major Comments:

**Comment 1:**
The title is misleading because the temporal variation of BD is treated in only one sub-chapter (Section 3.5), therefore I suggest to modify the title accordingly.
**Author Response:** We agree that the original title may have overemphasized the temporal aspect of the study. We have revised the title to better reflect the broader focus of the manuscript, which includes model benchmarking, feature importance, and temporal transferability. We propose the revised title *"Estimating changes in soil bulk density in Thailand using machine learning and remote sensing: model performance, feature importance, and temporal transferability"*.

**Comment 2:**

The contradiction in the application of the ANN model for 2009 Predictions is the most significant concern. The authors establish that the Artificial Neural Network (ANN) model is superior to other models, including tree-based methods like Random Forest and XGBoost. The feature importance analysis (Section 3.4, Figure 6) is key to this conclusion, showing that the ANN model uses a balanced set of predictors (slope, temperature, vegetation indices, etc.) and does not overly rely on Organic Carbon (OC). This is presented as a major strength, making the model more robust and generalizable. However, when applying this model to the 2009 dataset for temporal analysis, the authors state: "*...utilizing only OC data as the sole predictor for BD, as no ground-truth BD measurements were available for validation in that year*". This is a critical methodological flaw. The validated high-performance ANN model is a multivariate model that relies on a suite of remote sensing, topographic, and climate inputs. It cannot be applied using only a single input variable (OC). The authors need to clarify precisely how the 2009 predictions were made. Did they train a new, univariate ANN model using only OC? If so, its performance is unknown and unvalidated, and it cannot be claimed to be the "best-performing model.". Did they apply the original multivariate model but feed it only OC data, with placeholder values (e.g., zero, mean) for all other inputs? This would be invalid and produce meaningless results. As it stands, the entire temporal analysis (Section 3.5), including the reported 7.27% increase in mean BD and the 41.23% reduction in standard deviation, is not supported by the methodology. The conclusions about increased soil compaction and reduced variability are therefore unsubstantiated. The authors must either provide the full suite of predictor variables for 2009 and re-run the analysis or retract the temporal claims.

**Author Response:** Our wording created ambiguity about the 2009 prediction. To clarify, the 2009 projections used the full multivariate predictor set, processed identically to 2004 (remote-sensing indices, topography, climate). The 2004-trained ANN (frozen weights) was applied to the 2009 stack; no retraining, placeholders, or constant imputations were used. We have revised Section 2.7 accordingly.

*2.7 Model Implementations and Comparative Analysis*

*We trained six machine-learning models on the 2004 dataset, tuned hyperparameters via Bayesian Optimization, and selected the best model using RMSE, MAE, and $R^2$, verifying interpretability with permutation feature importance and partial-dependence style summaries to confirm balanced use of predictors. To assess portability across years, we constructed a 2009 multivariate predictor stack using the same variables and preprocessing as in 2004 (static terrain covariates reused; climate summaries and remote-sensing indices recomputed with identical definitions and compositing) and standardized all 2009 predictors with 2004 training statistics to prevent leakage; the frozen 2004-trained ANN was then applied to generate out-of-sample BD projections, produced only where the full predictor set was available. To examine year-specific sensitivity, we report permutation feature importance on the 2009 projections (as model-based sensitivity, not validation). For context, we also estimated 2009 BD using the top-performing PTFs from 2004 where inputs were available and compared distributional statistics (mean, SD, CV) between the PTF estimates and the ANN*

*projections, noting that 2009 field BD data were unavailable for independent validation.*

**Comment 3:**
Equation (1) on page 3 for calculating bulk density. The multiplication by 100 is incorrect. Soil bulk density is a measure of mass per unit volume, with standard units of g $cm^{-3}$. Multiplying by 100 would make the values physically meaningless (e.g., the reported mean of 1.28 g $cm^{-3}$ would become 128). This appears to be a significant typo that should be corrected. Please verify if this error propagated into any calculations or if it is merely a display error in the formula.

**Author Response:** We note for completeness that, following Reviewer 1's recommendation, we have removed Equations (1) and (2) from the manuscript. The previously shown "×100" in former Equation (1) was a typesetting error; all computations used correct units (g $cm^{-3}$) and were unaffected. After removal, the remaining equation (formerly Equation (3)) has been retained and renumbered for consistency. This also addresses the units concern raised in your comment. The corrected definition is:

$$\mathrm{BD} = \frac{\text{dry soil mass (g)}}{\text{volume of cylinder (cm}^3\text{)}} \qquad (1)$$

## Minor Comments:

**Comment 1:** Provide more information on satellite images: how many satellite products did you use in the data analysis after pre-processing? What's the satellite overpass frequency?

**Author Response:** We have added a dedicated paragraph in Section 2.2.1 *Landsat 5 Thematic Mapper (TM) and Pre-processing* detailing sensor characteristics, scene counts, and revisit as follows:

*2.2.1 Landsat 5 Thematic Mapper (TM) and Pre-processing*

*We used Landsat 5 Thematic Mapper (TM) Level-2 surface reflectance imagery for 2004 and 2009 (nominal 16-day revisit; 30 m spatial resolution; 6 spectral bands in the VIS, NIR, and SWIR). Data were accessed via the Google Earth Engine catalog. Given the study area's persistent tropical cloud cover, we ingested the full annual archive (1 January–30 December) and applied standard QA/CFMask-based cloud–shadow masking and basic BRDF/topographic normalization to ensure consistent inputs. The annual inventories comprised 769 scenes (2004) and 834 scenes (2009) prior to quality filtering. We then generated annual median composites (and derived spectral indices used as predictors) to mitigate residual clouds and temporal noise, yielding stable inputs for model training (2004) and out-of-sample projection (2009).*

**Comment 2:** In Section 2.8, temporal uncertainty is quantified as the absolute difference in standard deviations between the two years ($U=|\sigma 2009-\sigma 2004|$). While this measures the change in the *variability* or *dispersion* of BD predictions, it is not a standard definition of model uncertainty (which typically refers to prediction intervals or confidence in the estimates). The authors should consider rephrasing this to "change in spatial variability" to avoid confusion with predictive uncertainty.

**Author** Agreed. We have rephrased "*temporal uncertainty*" to "*change in spatial variability*" throughout Section 2.8, as recommended.

**Comment** 3**:** The correlation matrix (Figure 2) shows an exceptionally strong negative correlation between OC and BD ($r=-0.92$). This suggests that OC explains over 84% ($R^2 \approx 0.85$) of the variance in BD in the 2004 dataset by itself. This may limit the generalizability of the findings to regions where this relationship is less dominant. It would be beneficial for the authors to briefly discuss this in the context of their dataset and how it might influence model performance comparisons.

**Author Response:** We agree. We added a brief dataset-level note in Section 3.1 clarifying that the 2004 predictive signal is strongly driven by OC and that the results should be interpreted as a benchmark for that year's covariate pattern. We also expanded Section 4.2 to explain how this dominance can influence model comparisons, tree-based models tend to rely more on OC, whereas the ANN distributes importance more evenly, thus part of the ranking is contingent on the 2004 covariate structure.

> 3.1 Descriptive Statistics
> *"…The 2004 data exhibit a dominant OC–BD coupling ($r = -0.92$; $R^2 \approx 0.85$), while all other predictors show weak correlations ($r \leq 0.19$). This pattern indicates that much of the predictive signal in 2004 is attributable to OC within this dataset's covariate structure (climate, topography, and spectral indices for that year). Consequently, the 2004 results should be interpreted as a benchmark specific to this covariate pattern, rather than as evidence that OC will dominate to the same extent in other regions or years."*
>
> 4.2 Sensitivity of ML Models to Input Variables and Their Influence on Predicted BD Changes Over Time
>
> *"…Given the very strong OC–BD coupling in the 2004 dataset (Figure 2), part of the in-year accuracy of these OC-dominated models may reflect this dataset-specific relationship rather than a broadly generalizable pattern. In contrast, the ANN model demonstrated a more balanced use of predictors and maintained similar feature-importance behavior when applied to the 2009 dataset (Table 8), suggesting that it can adapt well to interannual variations. This finding indicates that the ANN model's predictive signal is not dominated by a single covariate, supporting its relative stability and potential transferability across different*

*temporal contexts, although we interpret 2009 results as model-based projections pending ground validation."*

**Comment 4:** Fix an objective evaluation for the RMSE obtained from different models. De Vos (2005) established satisfactory prediction performance for RMSE less than 0.25 g cm-3 (Palladino et al., 2022)

**Author Response:** We have added this benchmark in Section 3.2.

**Comment 5:** Figures: In caption for Fig. 1b, please clarify what the color bar is referring to. I see the red circles in the texture triangle simply indicate the soil samples. Then I see orange circles and colored texture classes. This is confusing. In Fig. 2 the authors should add the color bar title ("correlation"). The caption for Figure 4 reads "Loss function curves for neural network regression models... and learning curves for other machine learning models...". However, Figure 4 only shows these curves. The scatterplots are in Figure 5. The caption for Figure 4 should be corrected to only describe its own content. Increase font size of the axis tick labels, add a grid. In Fig. 5 increase font size of the axis tick labels, add a grid

**Author Response:** OK, we will fix this in the revised version.

**Comment 6:** Tables: In Table 2 replace "TPFs" with "PTFs". In Table 3, please add the coefficient of variation (CV). In Table 6, please specify how many data were used for calibration (training) and how many for validation (testing)

**Author Response:** Yes, we edited in Table 2, added coefficient of variation (CV) in Table3, and added detail about training (189 soil samples) and testing (47 soil samples) dataset in Table 6.

**Comment 7:** Inconsistent Units in Text: The manuscript occasionally uses inconsistent formatting for units. For example, in the abstract, the RMSE is given as "0.017 g cm³" , while in the ML performance table (Table 6), the MAE for the ANN model is listed as "0.012" without clear units in the text, and elsewhere as "0.012 g cm→". Please ensure consistent formatting (g $cm^{-3}$) throughout the manuscript for clarity and professionalism.

**Author Response:** Thank you, we will fix this in the revised version.

## Recommendation:

I recommend **Major Revisions**.
The manuscript addresses an important research topic and employs a robust and extensive comparative framework. The rigorous hyperparameter tuning and the scale of the PTF benchmark are significant strengths. However, the foundational contradiction in the

methodology for the 2009 temporal analysis is a critical flaw that invalidates the paper's primary conclusions regarding temporal trends in soil bulk density.
The authors must resolve this issue by either providing a valid methodology for the 2009 predictions using the full multivariate ANN model or by reframing the paper to focus solely on the 2004 model comparison, removing the temporal analysis. Additionally, the error in the BD formula and other minor points should be addressed. If the authors can satisfactorily resolve these major concerns, the revised manuscript would likely my opinion be suitable for publication.

REFERENCES

De Vos, B., Van Meirvenne, M., Quataert, P., Deckers, J., Muys, B., 2005. Predictive quality of pedotransfer functions for estimating bulk density of forest soils. Soil Sci. Soc. Am. J. 69 (2), 500–510.
Palladino M., N. Romano, E. Pasolli, P. Nasta. 2022. Developing pedotransfer functions for predicting soil bulk density in Campania Region. Geoderma 412, 115726 https://doi.org/10.1016/j.geoderma.2022.115726
**Citation**: https://doi.org/10.5194/egusphere-2025-2360-RC2

**Author Response:** We would like to clarify that our analysis for 2009 was conducted using the same full multivariate ANN model developed for 2004, incorporating all predictor variables (remote sensing indices, topographic, climatic, and soil covariates). However, the predictor datasets were recomputed using the 2009 environmental and remote sensing conditions, rather than reusing 2004 data. We have addressed this point explicitly in the Major Comments section for clarity. In addition, we have carefully addressed all minor comments to improve the manuscript's clarity and scientific rigor.

We hope that we have been able to satisfactorily address the issues that have been raised above from the reviewers. We look forward to hearing from you.

Sincerely,
Sunantha Ousaha
On behalf of all co-authors
10 October 2025