CC1: 'Comment on essd-2024-588', Tingxuan Zhang, 20 Mar 2025

(1) Random forest is a nonlinear supervised discrete classification model, while Pearson correlation coefficient is a correlation coefficient that measures linear correlation between two sets of data. The authors know nothing about it. They used the Pearson correlation coefficient to determine the variables for the random forest model inputs.

Thank you for your comment (1) regarding our choice of the Random Forest (RF) model and the use of the Pearson correlation coefficient for variable selection. We truly appreciate you raising this point, which allows us to clarify our methodology.

First, regarding the type of Random Forest model, we would like to clarify that Random Forest is a powerful ensemble learning algorithm capable of performing both classification and regression tasks. In this study, we specifically utilized Random Forest for regression prediction to estimate continuous Soil Organic Carbon Density (SOCD) values, which is a standard application of the algorithm in environmental science.

Second, concerning the use of the Pearson correlation coefficient to determine input variables for a nonlinear model like Random Forest, we have significantly refined our feature selection approach in response to valuable reviewer feedback. We now employ an enhanced three-stage feature selection methodology, which involves: ① Initial Screening: We first conduct a preliminary screening through Pearson correlation analysis ($p < 0.05$ significance threshold) to efficiently identify variables with potential linear relationships to SOCD and to remove overtly irrelevant features. ②Nonlinear Relationship Assessment: Subsequently, we incorporate Random Forest-based importance ranking to evaluate the nonlinear contributions of variables, ensuring that complex relationships are captured. ③Optimal Combination Optimization: Finally, we perform an exhaustive combinatorial optimization to select the variable combination that maximizes predictive performance (measured by R^2), while also applying a stricter multicollinearity threshold ($|r| > 0.95$).

Through this refined, multi-stage approach—which integrates statistical correlation analysis with machine learning-based feature importance assessment—we ensured that the final selected feature set (including 'temperature', 'elevation', 'NDVI', 'clay content', 'SR', 'BSI', and 'slope') comprehensively and optimally captures the key climatic, topographic, vegetation, and soil attributes influencing SOCD, while maintaining ecological interpretability.

These methodological refinements have been comprehensively incorporated into Section 3.2 ("Feature Selection Methodology") of the revised manuscript, and Figure 2 has been updated to visually present the final selected features and their relative importance. We believe this detailed

explanation clarifies our methodological choices and fully addresses your concerns.

(2) The authors claimed that they used climate zoning to improve the prediction, but this is absolutely unnecessary because temperature and precipitation are the two most important variables (shown in Figure 8) and are highly correlated to climates. In this case, why take the trouble of building the random forest model for each climate zone?

We greatly appreciate your insightful comments regarding the inclusion of climate zoning in our Random Forest (RF) model (Reviewer #2, Point 2 and Reviewer #3, Point 5), especially considering that temperature and precipitation were initially regarded as direct predictors. We understand the concern that this might introduce redundancy or suggest a limitation in the model's ability to independently capture complex relationships. We would like to clarify our approach and the subsequent refinements made to our feature selection process.

Updated Feature Set and the Role of Climate Zoning: Initially, our model considered 12 environmental factors, including temperature and precipitation. However, through an improved and rigorous feature selection method, we have refined our optimal feature set to comprise seven variables: mean annual temperature (Tem), elevation, NDVI, clay content (Clay), simple ratio index (SR), bare soil index (BSI), and slope.

It is important to note that only mean annual temperature (one of the parameters used to define climate zones) is now directly included as a predictor in the final seven-variable set used for our refined RF model. Precipitation is no longer a direct feature in the RF model.

Therefore, climate zoning, as described in Section 3.3, serves a distinct and crucial purpose beyond merely replicating direct predictors. Its primary role is: ① Geographical Stratification Strategy: Climate zoning is implemented to quantify the broad differences in temperature and precipitation across China. Given the vast and diverse environmental conditions across China, as highlighted by studies such as Tang et al. (2018), Soil Organic Carbon Density (SOCD) exhibits significant variations across different climatic zones. By categorizing China into four subzones (humid, semi-humid, semi-arid, and arid) based on multi-annual average temperature and precipitation thresholds (Mean Annual Precipitation MAP $\geq$ 400 mm, Mean Annual Temperature MAT $\geq$ 10°C), we are essentially performing a geographical stratification. ② Development of Zonal Models for Improved Accuracy: Within these distinct climatic subzones, we develop separate, localized SOCD estimation models. This strategy allows our models to better capture the unique environmental controls on SOCD within each zone. Relationships between environmental factors and SOCD can vary significantly across different climate types; by stratifying the data, our RF model can learn these unique, localized patterns with greater precision than a single, global model might, thereby significantly improving the predictive accuracy of SOCD estimates at a

regional scale.

Considerations on Model Efficiency and Deep Learning: While we acknowledge the theoretical point that highly efficient deep learning or AI foundation models might inherently learn these relationships without explicit geographic stratification, for our current Random Forest modeling approach, this explicit zoning provides a robust way to manage the inherent heterogeneity of the study area and ensures more accurate and reliable SOCD predictions. It serves a strategic purpose beyond mere training efficiency, enhancing the model's ability to capture regional nuances.

In conclusion, we believe that this refined approach—where only temperature is a shared variable between the climate zoning definition and the refined predictor set, coupled with the rationale for developing zonal models—provides a robust and justifiable methodology for our study, effectively improving the accuracy of SOCD estimation.

(3) Issue 2 leads me to my next big concern: the verification part. As the authors insist that climate zoning is the novelty of their methods, why did they verify their results against others across different climate zones? From a scientific view, climate zoning does not mean anything to improve the model's accuracy. For instance, if the authors did not use climate zoning but other geographical partitioning, the smaller the partition area, the more accurate the model would be considering Tobler's first law of geography states that everything is related to everything else, but near things are more related to each other.

Thank you for your concern regarding the verification part and the necessity of climate zoning in our methodology (Point 3). We understand your questions and are pleased to clarify our validation strategy and the crucial role of climate zoning in our study.

Firstly, regarding your question about "why did they verify their results against others across different climate zones," we would like to clarify that our validation approach was not simply comparing model results from one climate zone with observed data from another. Instead, we employed a **stratified spatial K-fold cross-validation method** for separating our training and testing samples, complemented by **independent measured sample validation, spatial validation, and temporal validation.**

Specifically, our validation strategy ensures:

1. **Spatial Representativeness:** The entire study area was divided into K (e.g., K=10) spatially distinct and non-overlapping sub-regions. The model was trained on K-1 sub-regions and independently validated on the remaining single sub-region. This ensures that we assess the model's generalization ability to geographically "unseen" areas, avoiding overestimation of model performance due to spatial autocorrelation.

2. **Temporal Representativeness:** Within each spatial fold, we ensured that samples from all three decadal periods (1980s, 2000s, 2010s) were proportionally represented, covering the entire observed historical span of data characteristics. This means our model's performance was rigorously evaluated within the full range of historical conditions represented by our dataset, and we are not performing any unvalidated temporal extrapolation beyond the broad historical windows from which our samples were drawn.

Secondly, regarding your assertion that "from a scientific view, climate zoning does not mean anything to improve the model's accuracy," we respectfully contend that **climate zoning is highly meaningful for improving regional model prediction accuracy.** As we detailed in our responses

to Reviewers #2 and #3, climate zoning serves as a **geographical stratification strategy** with the core objectives of:

- **Capturing Macro-Climatic Differences:** China is vast, and its climate zones (humid, semi-humid, semi-arid, arid) exhibit significant differences in temperature and precipitation regimes. These macro-climatic factors profoundly influence the distribution and accumulation mechanisms of SOCD.
- **Developing Regionally Tailored Models:** By segmenting the study area into climatically homogeneous subzones and developing separate, localized SOCD estimation models within each, we can better capture the unique environmental controls on SOCD in those specific regions. Relationships between environmental factors and SOCD can vary significantly across different climate types; by stratifying the data, our RF model can learn these unique, localized patterns with greater precision than a single, global model might, thus significantly improving the predictive accuracy of SOCD estimates at a regional scale.

Finally, concerning the hypothesis that "the smaller the partition area, the more accurate the model would be" (referring to Tobler's first law), while Tobler's law emphasizes that near things are more related, practical modeling processes are subject to significant constraints. Overly granular geographical partitioning can lead to:

- **Insufficient Sample Points:** Very small partitions might not contain a sufficient number of training sample points, thereby affecting the model's training quality and generalization ability. Model accuracy depends not only on regional homogeneity but also on the adequacy and representativeness of training data.
- **Computational Costs:** Building and running an extremely large number of models for very small areas would impose a significant computational burden.
- **Heterogeneity Challenges:** Even within very small regions, soil carbon distribution can exhibit substantial heterogeneity. Overly fine partitioning might not fully eliminate this heterogeneity, and could even reduce model stability due to sample size limitations.

Therefore, our climate zoning approach provides a strategy that balances regional homogeneity with practical modeling considerations (including sample point distribution and model efficiency), aiming for optimal overall predictive performance. We believe this combination of a robust validation strategy and an ecologically-based geographical stratification provides reliable and interpretable results for SOCD estimation.

(4) Even so, I do not find a significant improvement in the SOCD prediction compared to other published datasets.

Thank you for raising this critical point (Point 4) regarding the perceived lack of significant

improvement in our SOCD prediction compared to other published datasets. We understand your rigorous scrutiny of model performance and wish to provide a more detailed explanation and evidence to clarify the advancements and unique contributions of our study.

In fact, we believe that our study has achieved substantial progress in SOCD prediction, particularly in the following key aspects:

1. Significant Advantage of Climate Zoning Models over Global Models:

The core innovation of our method lies in the introduction of zoning models based on climate regions. As stated in the abstract of our manuscript, our zoning model demonstrably outperformed the global model without climate zoning in predicting SOCD. This improvement is quantitatively reflected in various model performance metrics (e.g., our model consistently showed better R^2, RMSE, and MAE values across different climate zones compared to the non-zoning global model). This indicates that for a region as vast and environmentally heterogeneous as China, building localized models through geographical stratification is more effective in capturing region-specific environmental-SOCD relationships, thereby significantly enhancing regional prediction accuracy and model robustness.

2. Filling a Gap for Long-Term, High-Resolution SOCD Products in China:

Existing published global SOCD products (such as SoilGrids250m, GSOCmap, HWSD v2.0, as shown in the comparisons in Figure 11 of the manuscript), while valuable, are typically developed at a global scale and may not fully capture the specific complexities and nuances inherent to the Chinese region. More importantly, there is a critical lack of high-spatial resolution (1 km) and long-time series (1985-2020) SOCD dynamic change datasets specifically for China. Our study aims to fill this crucial gap by providing continuous, fine-resolution spatial distribution and temporal dynamics of SOCD over nearly four decades. This long-time series product is essential for assessing the impacts of climate change and human activities on soil carbon pools, offering unique value that many existing static or short-term datasets cannot provide.

3. Comparative Analysis with Existing Datasets:

We conducted in-depth analyses comparing our results with several published datasets (including global products and the SOC Dynamics ML dataset for China, as shown in Figures 11 and 12). While direct "superiority" comparisons can be complex due to differences in methodologies, input variables, spatial/temporal scales, and target regions, our results demonstrate that:

Overall Performance is Competitive: Our SOCD prediction results are **highly comparable in accuracy metrics to these internationally renowned products, and in some aspects, demonstrate superior regional adaptability.** Particularly in China's complex and diverse geographical environment, a product specifically designed for this region, incorporating

refined feature selection and zonal modeling, often exhibits stronger local accuracy and ability to reflect regional heterogeneity.

Ability to Capture Regional Heterogeneity: Our zonal models are better able to reflect the true differences and spatial patterns of SOCD across China's distinct climate zones, which might be averaged out or smoothed in a single global model.

4. Rigorous Validation Strategy:

To ensure the rigor of our model evaluation and the reliability of our results, we employed a stratified spatial K-fold cross-validation and incorporated temporal stratification, ensuring the representativeness of both training and testing samples across space and time. This comprehensive validation strategy, coupled with independent measured sample validation, significantly enhances the credibility of our results, indicating robust model predictive capabilities under various conditions.

In conclusion, we believe that this study represents a significant improvement in terms of methodological innovation (climate zoning models), data product (long-time series, high-resolution SOCD dynamics product for China), and model performance (outperforming global models and being competitive with advanced products). We will further elaborate on these improvements in the discussion section of the revised manuscript and emphasize the unique value and contribution of our data product to future research.

(5) The highly skewed SOCD sample input leads to the model's low accuracy (Figure 4). This is probably one of many reasons why the accuracy of 0-20cm SOCD showed higher R2 than that of 0-100cm SOCD.

Thank you for your valuable observation regarding the potentially highly skewed SOCD sample input and its suggested link to model accuracy, as well as the difference in R^2 between 0-20 cm and 0-100 cm SOCD (Point 5). We greatly appreciate this insight and would like to elaborate on our understanding and approach.

We fully concur with your observation that **Soil Organic Carbon Density (SOCD) data, both globally and regionally, typically exhibits a skewed distribution (often right-skewed), which is also evident in our sample data (as shown in Figure 4).** This skewed distribution is a natural characteristic of soil carbon sequestration processes and can indeed pose challenges for certain aspects of modeling, particularly for predicting extreme values.

Regarding the concern that skewed data might lead to lower model accuracy, we would like to clarify:

1. **Robustness of Random Forest to Skewed Data:** As a non-parametric ensemble learning algorithm, Random Forest makes fewer assumptions about the distribution of input data and is

thus **inherently robust to skewed data.** It can effectively handle non-normal distributions and nonlinear relationships through the aggregation of decision trees.

2. **Data Transformation:** To further optimize model performance and mitigate the effects of skewed distribution, we applied a **log transformation** to the SOCD target variable during the modeling process. This standard data preprocessing technique effectively transforms skewed data to a more approximately normal distribution, which helps the model better capture relationships between variables and improves prediction stability and accuracy.

Concerning the higher R^2 for 0-20 cm SOCD compared to 0-100 cm SOCD, while sample data skewness might be a minor contributing factor, we believe the more fundamental scientific reasons are:

1. **Stronger Link of Topsoil SOC to Surface Environmental Factors:** SOC dynamics in the 0-20 cm depth (topsoil) are more directly and strongly linked to surface environmental factors such as climate, vegetation, topography, and human activities. These factors can be effectively acquired and quantified through remote sensing and meteorological data, allowing the model to better capture their driving mechanisms.

2. **Complexity and Inaccessibility of Deeper SOC Influencing Factors:** In contrast, SOC at 0-100 cm (deeper soil) is influenced by more complex, long-term, and less observable biogeochemical processes, such as slower decomposition rates, parent material characteristics, subsurface hydrology, and deeper root activity. These influencing factors are often less directly or reliably quantifiable and predictable using macro-scale or conventional remote sensing-derived covariates, leading to relatively weaker explanatory power from the model inputs.

3. **Sparsity and Uncertainty of Deeper Sample Data:** Measured data for deeper soil profiles are typically sparser in quantity and may have higher inherent measurement uncertainty or spatial variability compared to topsoil samples. Such data limitations also directly impact the accuracy of model predictions for deeper soil.

In summary, while the skewed distribution of SOCD samples is an inherent data characteristic, we have mitigated its impact through data transformation and by leveraging the robustness of the Random Forest model. The higher R^2 observed for 0-20 cm SOCD is primarily attributed to the stronger association between topsoil SOC and readily observable environmental factors, coupled with the inherent challenges in modeling deeper SOC, rather than simply being a consequence of sample skewness. We will clarify these explanations in the revised manuscript.

(6) Another reason is the adequate model input data. The lack of lidar data for soil depth measurement makes your results underestimated compared to other datasets (Figures 11 & 12).

Thank you for your concern regarding the adequacy of our model input data, specifically

your point that the lack of lidar data for soil depth measurement might lead to our results being underestimated (Point 6). We understand your focus on data quality and model accuracy, and we would like to provide a detailed clarification.

Firstly, regarding your assertion about the "lack of lidar data for soil depth measurement," we would like to clarify that **Lidar (Light Detection and Ranging) data is primarily used to acquire high-resolution surface topographic information (e.g., Digital Elevation Models, DEMs) and vegetation canopy structure data. However, it is generally not directly used for the direct measurement of soil depth (such as soil organic carbon profile depth) or the direct inversion of soil properties.** Soil depth is typically obtained through field boreholes, soil profile observations, or digital soil mapping approaches that infer soil properties based on covariates like topography and geology. While high-resolution topographic data can serve as an indirect auxiliary factor for soil spatial distribution, lidar data itself is not a necessary or primary input for direct soil depth measurement or SOCD inversion. Given the 1 km spatial resolution and long-time series (1985-2020) coverage of this study, nation-wide, long-term lidar data for soil depth measurement is currently not feasible to acquire and is not a conventional direct input variable in digital soil mapping.

Secondly, we firmly believe that **our model input data is sufficient and diverse, covering multiple key aspects necessary for SOCD modeling**, rather than being inadequate. We comprehensively utilized various authoritative and high-quality data sources, including:

- **Climatic Factors:** Long-term average temperature and precipitation data derived from meteorological stations, which are primary drivers affecting SOC accumulation and decomposition.
- **Topographic Attributes:** Elevation, slope, aspect, which control hydrothermal redistribution and soil erosion, significantly influencing SOC spatial distribution.
- **Vegetation Indices:** NDVI, Simple Ratio Index (SR), Bare Soil Index (BSI), etc., derived from Landsat satellite imagery. Vegetation is the main source of soil organic matter, and these indices effectively reflect vegetation cover and growth status, thereby characterizing their contribution to SOC.
- **Basic Soil Properties:** Clay and sand content, which are crucial indicators of soil texture, directly impacting soil physical structure and carbon sequestration capacity.
- **Measured SOCD Data:** A large volume of measured soil profile data used for model training and validation.

This comprehensive set of input data encompasses multiple dimensions including climate, topography, vegetation, and intrinsic soil properties, fully complying with the input data requirements of current mainstream Digital Soil Mapping (DSM) practices, and is sufficient to

support accurate SOCD prediction. Through our refined three-stage feature selection method, we ultimately identified seven optimal variables that effectively capture the key drivers of SOCD.

Finally, regarding your assertion that our "results are underestimated compared to other datasets" (Figures 11 & 12), we would like to emphasize:

- **Complexity of Comparisons:** Directly comparing data accuracy across different studies or products has inherent complexities. The purpose, input data sources, modeling methods, spatial and temporal resolutions, baseline years, and validation datasets used by different products can all vary. Therefore, judging "underestimation" based solely on visual impression without a unified, independent validation benchmark may not be accurate.
- **Regional Specificity:** Our study focuses on the Chinese region and innovatively employs a **climate zoning model**, aiming to capture the complexities and heterogeneity specific to China more precisely. This means our model might show different results in certain regional details compared to global models, but this difference often arises from our more refined capture of regional characteristics. For instance, the comparisons with other datasets in Figures 11 and 12 demonstrate consistency in spatial distribution patterns and localized differences, which does not imply underestimation, but rather reflects the varied performance of different methods and input data in specific regions.
- **Internal Validation Results:** Most importantly, we achieved competitive model accuracy metrics through rigorous internal validation strategies, including **stratified spatial K-fold cross-validation** and **temporal stratification validation.** These quantitative validation results (e.g., R^2, RMSE, MAE) fully demonstrate the robustness and reliability of our model, proving its effectiveness in predicting SOCD across China.

In conclusion, we believe that even without relying on lidar data for soil depth measurement, our model input data is sufficiently comprehensive and robust. Furthermore, through our innovative climate zoning methodology and stringent validation process, our SOCD prediction results are reliable and hold unique value, especially in terms of long-time series and regional refinement.

(7) Figures 5(a) and 5(c) are unnecessary as the authors did not conduct any analysis using the biomes.

Thank you for your comment regarding Figures 5(a) and 5(c), suggesting they might be unnecessary as we did not conduct any direct analysis using biomes (Point 7). We understand your consideration and would like to clarify the intended purpose of these figures.

We agree that biomes were not directly used as predictor variables or as independent modeling zones in our final Random Forest model. However, the purpose of Figures 5(a) and 5(c) is to provide readers with **crucial ecological background and information on the environmental heterogeneity of the study area (China).**

Specifically, these figures serve to:

1. **Provide Macro-Ecological Context:** China is a vast country encompassing a wide variety of ecosystem types. Figures 5(a) and 5(c), by displaying the distribution of major biomes, help readers visually grasp the macro-ecological patterns of the study area. This background information is essential for understanding the distribution and variation of Soil Organic Carbon Density (SOCD) across different geographical regions, as it reflects the integrated outcome of long-term interactions between climate, vegetation, and soil.

2. **Support the Rationale for Climate Zoning:** Although we did not directly use biomes for modeling, the demarcation of biomes itself is strongly influenced by climatic conditions (e.g., temperature and precipitation). By illustrating these biomes, we aim to further emphasize the immense heterogeneity of China's terrestrial ecosystems. This heterogeneity precisely underpins our rationale for adopting **climate zoning for geographical stratification in our modeling approach** (rather than a single global model). It visually reinforces the necessity of considering that SOCD's relationships with environmental factors might differ across distinct ecological-geographical regions.

3. **Enhance Readability and Comprehension:** For general readers or those less familiar with China's geographical environment, a straightforward biome map can quickly establish an understanding of the study area's complexity, thereby facilitating a better comprehension of the drivers of SOCD spatial distribution and the applicability of our research methodology.

In summary, Figures 5(a) and 5(c) are not direct inputs for model analysis but serve as **important background information and contextual descriptions.** Their inclusion aims to enhance the reader's understanding of China's ecological diversity and indirectly support the necessity of our regionalized modeling approach using climate zoning. We believe they contribute to the manuscript's readability and the completeness of its scientific context.