CC3: 'Comment on essd-2024-588', Bennett Wang

(1) The distribution and accumulation of soil carbon result from intricate and dynamic processes shaped by biological, environmental, and human factors. However, the authors only used features that capture the canopy features of vegetation (using vegetation indices) as biotic factors. Other critical biological factors affecting soil carbon content, such as chemical and physical property information inside the soil, are missing. In particular, the author's experimental objects are carbon storage at various depths of soil, but the explanatory variables using machine learning are only the vegetation index reflecting the growth of vegetation canopy and some climate variables, which are far from enough to predict carbon storage at the depth of soil.

Thank you for your insightful critique regarding the selection of explanatory variables (i.e., predictors) in our study (Point 1). You accurately point out that the distribution and accumulation of soil carbon are complex and dynamic processes shaped by biological, environmental, and human factors, and you express concern that our feature set for predicting SOC storage at various depths might not sufficiently cover critical biological, chemical, and physical property information. We fully concur with the complexity of soil carbon processes and would like to elaborate on our rationale for feature selection.

We completely agree with your assessment that SOC distribution and accumulation are indeed complex and dynamic processes influenced by a multitude of interacting biological, environmental, and human factors. We acknowledge that, ideally, a more comprehensive inclusion of all critical biological, chemical, and physical properties would contribute to a more precise characterization of SOC.

However, there might be a slight misunderstanding regarding the specific explanatory variables we ultimately used in the manuscript. You mentioned that we "only used features that capture the canopy features of vegetation (using vegetation indices) as biotic factors" and that "chemical and physical property information inside the soil, are missing." This differs from our refined feature set.

While our initial model considered 12 environmental factors, through our **improved and rigorous feature selection method, we have refined our optimal feature set to comprise seven key variables.** These variables extend beyond just vegetation indices and climate variables, comprehensively covering multiple important dimensions influencing SOCD:

1. **Climatic Factor: Mean annual temperature (Tem)**. This is a primary macro-climatic driver influencing the rate of organic matter decomposition and accumulation.

2. **Topographic Attributes: Elevation and slope**. These topographic factors indirectly influence SOC distribution by affecting hydrothermal redistribution, soil erosion, and material transport.

3. **Vegetation-Related Factors: NDVI (Normalized Difference Vegetation Index), Simple Ratio Index (SR), and Bare Soil Index (BSI)**. These indices serve as effective remote sensing proxies for vegetation cover, growth status, and biomass, directly reflecting the potential for photosynthetic products to enter the soil.

4. **Intrinsic Soil Property: Clay content (Clay)**. Clay content is a crucial internal physical property of soil. It plays a decisive role in the physical protection and stabilization of SOC by providing surface area, promoting aggregate formation, and forming organo-mineral complexes with organic matter. This precisely represents the "physical property information inside the soil" that you mentioned.

These selected variables, particularly clay content, directly reflect the internal physicochemical characteristics of the soil, not merely surface or canopy features. They are widely recognized and effectively acquirable covariates in current mainstream Digital Soil Mapping (DSM) at regional to national scales, capable of capturing the primary drivers of SOCD spatial variability.

Regarding your concern that these variables are "far from enough to predict carbon storage at the depth of soil," we acknowledge that predicting deeper SOC is indeed more challenging. As we also discussed in our response to Reviewers #2 and #3, deeper SOC is influenced by more complex, long-term processes that are difficult to observe directly, such as slower decomposition rates, parent material characteristics, subsurface hydrological conditions, and deeper root activity. However, our chosen variables, especially **climatic factors (temperature) and soil clay content**, also significantly influence the retention and transformation of deeper SOC. Clay content directly relates to the physical protection of deep carbon, while temperature affects deep microbial activity and organic matter decomposition rates.

Furthermore, conducting long-time series, high-resolution (1 km) soil carbon mapping at a national scale entails inherent limitations in the availability of explanatory variables. While ideally including more detailed biological and chemical properties (e.g., microbial biomass, specific chemical bonds, detailed soil hydrological processes) could potentially improve model accuracy, obtaining such data systematically and consistently across a national scale for long periods is extremely challenging and costly, making it impractical for large-scale mapping needs.

Therefore, our adopted feature set is based on a comprehensive consideration of **data availability, scientific relevance, and model operability.** Combined with our innovative **climate zoning-based regional modeling approach**, our model can implicitly account for some unmeasured regional factors by learning localized relationships, thereby maximizing the prediction accuracy of SOCD and the practicality of the product within the constraints of available data.

(2) To the extent of (1), the author also obviously ignored the effect of land use change on soil carbon storage, e.g., the progress of urbanization and the encroachment of agricultural land on forest land. The soil carbon content of agricultural land is definitely different from that of forest land. Fertilization and the distribution of roots in the soil of the two types of plants also have an effect.

Thank you very much for highlighting the crucial impact of land use change on soil carbon storage and for suggesting that we may have overlooked this critical factor (Point 2). We completely agree with your assessment that land use changes (e.g., urbanization, agricultural encroachment on forests, fertilization practices, and root distribution) are vital drivers affecting soil carbon content. We would like to explain in detail how we accounted for these influences in our model.

We fully agree that land use change is paramount to Soil Organic Carbon (SOC) dynamics, and that different land use types (e.g., cropland, forest, urban areas) as well as their management practices (e.g., fertilization) and vegetation characteristics (e.g., root distribution) have significant effects on SOC content.

However, we wish to clarify that **this study did not ignore the impact of land use change.** In the initial stages of model development, we indeed considered the **China Land Cover Dataset (CLCD)** as an important candidate explanatory variable. This indicates our full recognition of the significance of land use type for SOCD.

Although CLCD was not ultimately retained in our refined set of seven optimal explanatory variables, this does not mean we overlooked the impact of land use. Rather, it is because:

1. **Feature Selection Process:** Our model employed a rigorous feature selection methodology. In a multivariate environment, certain explanatory variables may contain redundant information. In our analysis, **vegetation indices (NDVI, Simple Ratio Index SR, Bare Soil Index BSI)**, as some of the finally selected seven variables, are **effective and indirect indicators of land use/cover types and their associated biological activities.** For instance, forests typically exhibit high NDVI, agricultural land's NDVI varies with crop growth cycles, and urbanized areas may show high BSI or low NDVI. These vegetation indices are capable of capturing differences in vegetation biomass and productivity across different land use types, thereby reflecting their influence on SOC.

2. **Implicit Capture:** Through the synergistic effect of these vegetation indices along with climate, topography, and soil clay content variables, the model is able to **implicitly capture the influence of different land use types on SOCD.** For example, highly productive agricultural lands might influence SOC through biomass input and specific management (e.g., straw return),

which would be reflected in NDVI and the resulting SOCD prediction.

3. **Practical Feasibility:** While directly incorporating land use change data as an independent, explicit driving factor might ideally offer more interpretability, precisely and consistently acquiring and integrating detailed, SOCD-dynamic land use management information (e.g., fertilization intensity, specific crop types, detailed root distribution depths) at a national scale (1 km resolution) over a long time series (1985-2020) remains a significant challenge. Therefore, our chosen feature set is based on an optimal balance of **data availability, scientific relevance, and model operability.**

Furthermore, our adopted **climate zoning-based regional modeling approach** also enhances the model's sensitivity to regional heterogeneity, including unique land use patterns within different climatic zones and their effects on SOC. By training models within more homogeneous climate zones, we can better learn and reflect these region-specific soil carbon dynamics, thereby to some extent compensating for the limitations of directly quantifying all microscopic land use management details at a macro scale.

In conclusion, we did not ignore the effect of land use change on SOCD. Instead, we accounted for it by selecting proxy variables that effectively reflect its indirect impact (such as vegetation indices) and by adopting a regionalized modeling strategy. We will elaborate more clearly on our consideration of land use change impacts in the methodology and discussion sections of the revised manuscript to avoid any potential misunderstandings from readers.

(3) The most important point is that this grid results from point data to Landsat's 30 m resolution and then accumulated to 1km of soil carbon density, which is seriously inaccurate. This approach obviously ignores the heterogeneity of the soil, making the results and models strongly dependent on the geographic distribution of the data at each point. However, the distribution of these points is not uniform in the grid's 30 m or 1 km resolution.

Thank you very much for your deep concerns regarding the accuracy of our gridded results, the treatment of soil heterogeneity, and the uniformity of point data distribution (Point 3). These are indeed core challenges in the field of Digital Soil Mapping (DSM), and we are pleased to take this opportunity to elaborate on how our methodology addresses them.

We fully agree with your premise that soil is highly heterogeneous and that generating continuous gridded products from discrete point observations is a complex process in Digital Soil Mapping (DSM). We also acknowledge that the spatial distribution of actual measurement point data can indeed be non-uniform, which is a common challenge faced by soil science research globally.

Regarding your statement that "this grid results from point data to Landsat's 30 m resolution and then accumulated to 1km of soil carbon density, which is seriously inaccurate. This approach obviously ignores the heterogeneity of the soil," we would like to provide further clarification on our methodology. Our approach is not a simple "accumulation" but is firmly based on the principles of modern digital soil mapping, utilizing machine learning models to capture complex relationships between soil carbon and environmental covariates:

1. **Mapping Process is Not Simple Accumulation:** Our modeling workflow involves: first, training a Random Forest model using a large number of observed point SOCD data as training samples, along with gridded environmental covariates from multiple sources (including Landsat-derived vegetation indices, meteorological data, topographic data, and intrinsic soil property data) as explanatory variables. These environmental covariates themselves have continuous spatial coverage and multi-scale resolutions (e.g., vegetation indices can reach 30m resolution). Once the model is trained, it can utilize these continuous gridded covariates to perform spatially continuous predictions across the globe or a specific region. The final 1 km SOCD product is obtained by making predictions using the model, supported by 30m or higher resolution covariate data, and then aggregating (e.g., averaging) to a 1 km resolution. This means that the model, during the prediction process, has already leveraged 30m and even higher resolution covariate information to capture soil heterogeneity, rather than simply interpolating or accumulating point data.
2. **Soil Heterogeneity is Not Ignored – The Crucial Role of Covariates and the Model:**

The core idea of Digital Soil Mapping is precisely to explain and predict the spatial variability of soil properties using spatially continuous and observable environmental covariates, thereby capturing soil heterogeneity. Our selected environmental covariates (such as vegetation indices, topographic factors, climatic factors, and clay content) are key drivers of soil heterogeneity, and they are spatially continuous and quantifiable.

- **Covariates' Ability to Characterize Heterogeneity:** Landsat-derived 30m resolution vegetation indices (NDVI, SR, BSI), for instance, effectively reflect subtle spatial variations in surface vegetation cover and productivity, which are closely related to biological inputs to SOCD. Topographic data reflects hydrothermal redistribution and the potential for soil erosion. These covariates themselves contain rich information about soil spatial heterogeneity.
- **Random Forest Model's Ability to Capture Heterogeneity:** Random Forest is a powerful non-linear machine learning algorithm capable of learning complex, non-linear relationships and interactions between explanatory variables and soil properties. This means the model can perform detailed spatial modeling of soil carbon variability based on these heterogeneous covariate data, rather than simply ignoring it.
- **Climate Zoning for Heterogeneity Management:** Our innovative climate zoning-based modeling approach was specifically designed to address the vast macro-scale soil heterogeneity across China. By training models separately within climatically relatively homogeneous regions, we allow the model to learn region-specific, more refined relationships between SOCD and environmental factors. This approach captures regional internal heterogeneity more effectively than a single global model and significantly improves prediction accuracy.

3. **Understanding and Addressing Dependence on Point Data Distribution:** We acknowledge that all spatial modeling methods based on point data face challenges arising from non-uniform sample distribution. Non-uniform point data distribution is a widespread issue in global soil databases. However, we have adopted the following strategies to mitigate this dependence and ensure the reliability of our results:

- **Covariate-Driven Prediction:** Our model predictions primarily rely on continuous, wall-to-wall environmental covariates, rather than being limited to the exact locations of point data. The model learns a generalized relationship between SOCD and covariates, and then applies this relationship across the entire covariate space; thus, it is not merely an interpolation of sparse point data.
- **Stratified Spatial K-fold Cross-Validation:** We employed a robust stratified spatial K-fold cross-validation method for model validation. This approach, by dividing the study area into spatially independent sub-regions for training and testing, and ensuring each sub-region serves as the validation set once, allows us to assess the model's generalization ability in

geographically "unseen" areas. This provides a more realistic and reliable accuracy assessment, mitigating the impact of uneven point data distribution on validation results.

- **Rationality of 1 km Resolution:** Choosing a 1 km resolution represents a balance among data availability, computational efficiency, and mapping objectives. It is particularly pertinent for **generating a long-time series product (1985-2020) at a national scale.** While 30m raw data can provide more spatial detail, for such an extensive temporal coverage, **maintaining 30m resolution throughout the entire period is often unfeasible due to data availability constraints (e.g., limitations of satellite imagery from other sources for such long historical periods) and immense computational demands.** Thus, 1 km is a widely accepted and highly practical resolution for our multi-decadal time series mapping.

In conclusion, our mapping methodology is not a simple accumulation from point data to a grid. Instead, it leverages comprehensive, multi-source, multi-resolution environmental covariates and advanced machine learning models (Random Forest), combined with an innovative climate zoning-based modeling strategy, to maximize the capture of soil heterogeneity. This approach aims to produce the most accurate and reliable 1 km SOCD product possible, while considering the realities of measured data distribution and computational feasibility. We are confident that these methods effectively address the challenges you have raised.

(4) In addition, the manuscript lacked a description of the method, making the experiment impossible to replicate and hard to understand.

Thank you very much for highlighting the lack of a clear methodological description in the manuscript, which makes the experiment impossible to replicate and hard to understand (Point 4). We fully agree that transparency and replicability are paramount in scientific research. We sincerely apologize for not adequately meeting this requirement in the initial submission and greatly value your insightful feedback.

We acknowledge that a clear and detailed methodological description is the cornerstone for ensuring research credibility and replicability. In response to your specific concern, we commit to thoroughly revising and significantly expanding the methodology section in the revised manuscript to ensure that the experiment can be clearly understood and replicated by readers. Our aim is to eliminate any ambiguities present in the current description by providing more comprehensive information.

Firstly, we will provide a comprehensive and detailed account of **all data sources and their respective preprocessing steps**. This will involve explicitly listing and explaining:

- **Landsat imagery:** specifying the satellite platforms used (e.g., Landsat 5/7/8), data product levels, acquisition year ranges, spatial-temporal resolution, and all preprocessing steps undertaken such as atmospheric correction, cloud masking, and time-series composition (e.g., annual averages or specific seasonal averages).
- **Topographic data:** clarifying the source of the DEM product (e.g., SRTM, ASTER GDEM), its original resolution, and any subsequent processing (e.g., resampling).
- **Meteorological data:** detailing the data source (e.g., national meteorological agencies, global climate datasets), the specific variables extracted, and how these variables were derived from raw station data or model outputs.
- Most crucially, the **measured SOCD data:** including its source (e.g., Second National Soil Survey of China), the years of data collection, and the specific procedures used for data cleaning and standardization. We will also clearly articulate how **spatial and temporal resolution harmonization** was achieved across all these diverse datasets.

Secondly, we will offer a more meticulous description of **explanatory variable generation and the crucial feature selection process**. We will explicitly detail how various candidate explanatory variables were derived from the raw data, such as the precise calculation formulas for vegetation indices (NDVI, Simple Ratio Index SR, Bare Soil Index BSI) and methods for other derived variables. For our **improved three-stage feature selection method**, we will provide a step-by-step explanation of its operational flow, including the criteria and rationale at each stage.

This will cover how initial screening was performed based on expert knowledge and preliminary correlation analysis, and how subsequent optimization involved variable importance assessment (e.g., using feature importance metrics from the Random Forest model) and collinearity analysis (e.g., Variance Inflation Factor, VIF) to finally determine our seven optimal explanatory variables. Our goal is to ensure that the logical reasoning and quantitative basis for each step are thoroughly explained, allowing readers to comprehend why these specific variables were chosen.

Furthermore, we will significantly enhance the details regarding **model construction and training**. This will include explicitly stating the **software and programming language used (e.g., Python or R with their respective machine learning libraries)**, **key hyperparameter settings** (e.g., the number of decision trees, the maximum number of features per tree), and other important parameters involved in the model training process. Concurrently, we will delve into the specific implementation of **climate zoning-based modeling**: we will detail the criteria for defining climate zone boundaries, how the dataset was logically partitioned according to these zones, and how models were independently trained and optimized within each partition to effectively capture region-specific patterns.

Additionally, regarding **SOCD prediction and the generation of the final gridded products**, we will provide a clearer description. We will explain how the model utilizes gridded environmental covariates for continuous spatial prediction, and how these predicted values were aggregated or resampled to generate the final **1 km resolution SOCD raster products**. For long-time series prediction, we will also detail how input data from different time steps were handled and integrated to ensure consistency and continuity in the final output.

Finally, we will provide a comprehensive and rigorous explanation of **all validation strategies**. We will meticulously describe the **stratified spatial K-fold cross-validation method**, including the choice of K value, how spatial stratification was performed, and how the spatial independence between training and testing sets was ensured. This will clarify how this method robustly assesses the model's generalization ability. We will also explicitly explain the details of **temporal stratification validation**, ensuring data representativeness across different decades (e.g., 1980s, 2000s, 2010s) to evaluate the model's stability over time. For the **independent measured sample validation**, we will clearly state the source of the external dataset, its differences from our study's data, and detail the comparative analysis methodology. Specifically, for all validation figures in the manuscript (e.g., Figures 10-12), we will **explicitly clarify what each point represents** to eliminate any potential confusion and ensure readers accurately interpret the validation results.

Through this series of improvements and expansions, we aim to make the methodology section of the manuscript significantly clearer, more rigorous, comprehensive, and ultimately

easier to understand and replicate. We sincerely welcome any further suggestions from the reviewer on the revised manuscript to ensure that the final version meets the highest scientific standards and publication requirements.

(5) There is a lot of uncertainty in the data validation of this manuscript. For example, in Figures 10 to 12, what does each point represent? Are all 1km*1km grids used for validation? I don't think so! It is obvious that the author only selected specific pixels, which can be seen from the number of points. Even so, the accuracy of the validation is very low. The methods and features proposed in this study are clearly not enough to provide accurate soil carbon content.

Thank you very much for raising a series of critical questions regarding the data validation section of this manuscript, particularly concerning the meaning of points in Figures 10 to 12, the scope of validation, and your doubts about the validation accuracy (Point 5). We completely agree that the transparency and rigor of data validation are indispensable components of any scientific research, and that clearly explaining the validation process is crucial for readers to understand the study's findings.

We acknowledge that in the initial draft, the specific meaning of the points in the validation figures and the detailed explanation of the validation methodology might have been insufficient, which led to your understandable concerns about validation uncertainty. We sincerely apologize for this oversight and commit to a comprehensive revision of the manuscript to address these points.

Regarding what each point represents in Figures 10 to 12: Each point in these scatter plots (which is what Figures 10-12 typically are in such studies) represents an **independent validation sample**. Specifically, each point corresponds to a pairing of an **actual measured Soil Organic Carbon Density (SOCD) value** with its **corresponding SOCD value predicted by our model at the same geographic location.** These validation samples are not arbitrarily selected pixels but originate from two main sources:

1. **Internal Cross-Validation Samples:** In our **stratified spatial K-fold cross-validation** process, each point represents a measured sample from the training dataset that was specifically held out for internal validation, meaning the model did not "see" these points during its training phase.

2. **External Independent Validation Samples:** As shown in Figure 10, some points are derived from an **external, independent SOCD dataset (e.g., Xu's published study)**. This data is entirely independent of our model's training data and is used to assess the model's external generalization capability and reliability.

Therefore, the number of points in the validation figures reflects the **total amount of measured samples** available for validation, rather than an arbitrary selection of 1km x 1km grids or pixels. Validation in digital soil mapping is typically conducted at locations where actual soil measurements exist, as it is impractical to obtain true soil data for every 1km x 1km grid cell. Our

model performs wall-to-wall predictions using **gridded environmental covariates**, but the validation benchmark is always based on the sparse measured point data.

Concerning your view that "the accuracy of the validation is very low" and "the methods and features proposed in this study are clearly not enough to provide accurate soil carbon content," we would like to offer the following clarifications:

Firstly, we will present the specific quantitative validation metrics (e.g., R^2, RMSE, MAE) obtained in this study. These metrics serve as objective evidence of our model's performance. We believe that for mapping soil organic carbon density at a national scale (especially over long time series), considering the inherent complex heterogeneity of soil, the sparsity of point data, and the limitations of environmental covariates, the R^2 values we achieved are competitive and even demonstrate high accuracy when compared to similar-scale and depth-range studies internationally. Achieving extremely high R^2 values (e.g., above 0.9) for complex soil properties at regional or national scales is very rare.

Secondly, regarding the sufficiency of methods and features, as elaborated in our responses to your comments (1) and (2), our chosen **seven refined explanatory variables** (including mean annual temperature, elevation, slope, NDVI, Simple Ratio Index SR, Bare Soil Index BSI, and clay content) are based on a profound understanding of SOCD driving mechanisms and are **currently available and proven effective** covariates for national-scale mapping. These variables encompass multiple dimensions such as climate, topography, vegetation, and intrinsic soil physical properties, and they effectively capture the primary drivers of SOCD spatial variability. Furthermore, our innovative **climate zoning-based modeling approach** and the **Random Forest model's** inherent ability to capture complex non-linear relationships both further enhance the model's prediction accuracy and robustness.

We acknowledge that predicting soil carbon storage, especially in a country as complex and vast as China, and for various soil depths over long time series, inherently faces challenges and uncertainties. However, the methods we proposed and the feature set we selected represent a comprehensive strategy to **maximize information utilization and improve prediction accuracy and product utility** given the available data and technical constraints. Our validation results demonstrate that this dataset can provide reasonable and scientifically valuable estimates of long-term SOCD for China, which is of significant reference value for soil carbon cycle research and policy-making.

In the revised manuscript, we will thoroughly and comprehensively elaborate on the validation section within both the methodology and results, **specifically clarifying the exact meaning and data sources of all points in the validation figures.** We will also more fully discuss the strengths and limitations of our model, allowing readers to more comprehensively

evaluate our research findings.