RC3: 'Comment on essd-2024-588', Anonymous Referee #3, 07 May

1) I don’t understand why the authors calculated and incorporated vegetation and water indexes as predictors instead of using the surface reflectances at some key bands as predictors of the Random Forest Model. As we know, machine learning models can learn the complicated and nolinear relationships.

We sincerely appreciate your insightful question regarding our choice of vegetation and water indices as predictors over raw surface reflectances in our Random Forest model. We understand that machine learning models are adept at learning complex, non-linear relationships directly from raw spectral bands, and your point raises an important methodological consideration.

Our decision to incorporate derived spectral indices (such as NDVI, SR, and BSI) instead of raw band reflectances was the result of an improved, three-stage feature selection method implemented in our refined analysis, aiming for optimal predictive performance and ecological relevance. While raw bands provide fundamental spectral information, indices are often designed to specifically highlight biophysical properties (e.g., vegetation density, soil moisture) that have a more direct and synthesized relationship with soil characteristics like SOC. Our rigorous selection process first involved a Pearson correlation analysis ($p<0.05$) to identify potential linear relationships. Following this, we employed Random Forest importance ranking to evaluate the non-linear contributions of various potential predictors, including both raw bands and derived indices. Crucially, in the final stage, we performed an exhaustive combination optimization to select the variable set that yielded the best predictive performance (highest R2), while also applying a stricter multicollinearity threshold ($|r|>0.8$). Through this comprehensive process, we found that derived indices like SR and BSI, alongside NDVI, offered superior predictive power and a more robust representation of key vegetation-soil interactions, outperforming the direct inclusion of raw bands which often exhibited higher inter-band correlation and potentially less direct ecological meaning for SOC estimation. This refined feature set, which now includes 'temperature', 'elevation', 'NDVI', 'clay content', 'SR', 'BSI', and 'slope', comprehensively captures climate, topography, vegetation, and soil attributes in a manner optimized for SOCD prediction.

We believe this integrated three-stage approach, which combines statistical correlation analysis with machine learning feature importance evaluation and explicit multicollinearity control, ensures the selection of the most meaningful and predictive variables while maintaining ecological interpretability. We trust this detailed explanation clarifies our methodological choices. Please let us know if you have further questions or require additional analysis.

2)The exact years when the soil samples were collected seems to be unknown. So, how did you select the corresponding annual predictors, such as the indexes derived from Landsat images? I doubt the lack of exact sample collection time could lead to uncertainty in the final annual SOC products, especially in terms of interannual variation.

Thank you for your insightful question regarding the uncertainty introduced by the unconfirmed exact soil sample collection years, particularly concerning the selection of annual predictors like Landsat-derived indices and its potential impact on interannual variation in our final SOC products. Your query highlights a critical challenge when working with historical soil data, and we've given this considerable thought.

First, it's important to consider the inherent stability of soil organic carbon (SOC). Unlike rapidly changing parameters such as surface soil moisture or vegetation indices, SOC, as a long-term carbon pool in the soil, typically doesn't exhibit significant variations over short periods (e.g., one year or a few years). Noticeable SOC changes usually require a longer timescale, often five years or more, to be detected. Based on this characteristic, our study isn't aimed at generating precise annual SOC products. Instead, we conduct our mapping in five-year periods (e.g., 1985-1990, 1990-1995, etc.). We consider SOC changes within these five-year spans to be relatively stable. Accordingly, we process annual predictors, such as Landsat-derived indices, into five-year average values to represent the mean conditions of that period. This approach ensures temporal consistency and representativeness, which helps to mitigate uncertainty stemming from imprecise sample collection times for interannual variation.

Second, you're correct that we can't provide the exact annual collection dates for all 8,203 sample points due to the nature of our data sources (large-scale national soil surveys and research projects). However, we've carefully categorized these samples into three significant historical decades based on their primary collection periods: the 1980s (from the 1980-1996 census data), the 2000s (from the 2000-2014 census data), and the 2010s (primarily from post-2010 national soil system records). These datasets, by themselves, cover broad timeframes within their respective decades.

To maximally ensure our model learns from and generalizes across different historical periods, especially given the lack of precise annual sample times, we employed a stratified spatial K-fold cross-validation method. This approach not only ensures that our training and testing sets are spatially independent but also incorporates temporal (decadal) stratification. We've made sure that in every cross-validation fold, samples from the 1980s, 2000s, and 2010s are proportionally represented. This means our model is consistently exposed to and validated against the soil characteristics and environmental conditions of all three observed decades. This strategy effectively enables the model to capture and reflect decadal-scale trends and spatial patterns of

SOCD within the observed timeframe, rather than attempting to resolve precise annual fluctuations.

In summary, by leveraging the inherent stability of SOC, generating our products in five-year periods, and employing a stratified spatiotemporal cross-validation strategy, we've minimized the impact of sample collection time uncertainty. Our products are designed to robustly reflect the decadal-scale changes and spatial distribution of SOCD in China, ensuring that our inferences are firmly grounded within the observed temporal and spatial ranges. We hope this detailed explanation addresses your concerns.

3)Line 145: You should highlight what are the potential shortcomings of SOC Dynamics ML dataset. Without detailed description on this, it's unclear why you would like to produce a similar dataset.

Thank you for your insightful question regarding the relationship between our dataset and the existing SOC Dynamics ML dataset, and for prompting us to clarify the unique contribution of our work. We appreciate the opportunity to elaborate on why producing a new, similar dataset is necessary.

We acknowledge the significant value of the SOC Dynamics ML dataset (Li et al., 2022) as a valuable resource for understanding SOC dynamics and its drivers in China. However, our study focuses on a distinct and critical aspect of soil carbon, which justifies the development of our new dataset.

The core distinction lies in our explicit emphasis on producing a Soil Organic Carbon Density (SOCD) dataset at a 1 km resolution for specific depths (20cm and 100cm) spanning 1985 to 2020. While the existing SOC Dynamics ML dataset provides valuable information including "SOC content" and "SOC stocks," our work specifically addresses the rigorous conversion from raw SOC content to SOCD. This conversion, as detailed in Section 3.1 of our manuscript, involves the precise integration of bulk density and coarse fragment data, which are crucial for accurate carbon accounting and inventorying at various scales. Many applications, particularly those focused on carbon budgeting and policy-making, require standardized SOCD values rather than just content or stocks calculated with varying methodologies.

Furthermore, our dataset offers unique advantages that complement existing resources:

(1) Standardized Density Metric: Our focus is on providing a consistent SOCD product, ensuring methodological transparency and comparability across different regions and depths, which might not be uniformly emphasized or detailed in other "SOC content" or "SOC stock" datasets.

(2) Temporal Specificity and Integration: We rigorously integrate historical soil survey data

(1980s, 2000s, 2010s) to map the spatial distribution of SOCD across these distinct periods, offering a snapshot of density changes over a specific long-term timeframe (1985-2020).

(3) Refined Methodological Approach: As discussed in previous responses, our work employs an improved three-stage feature selection method and a stratified spatial K-fold cross-validation strategy to ensure the robustness and spatial generalization capabilities of our SOCD estimates, leveraging diverse environmental predictors.

In essence, while existing datasets provide broad insights into SOC, our contribution is to provide a high-resolution (1km), precisely calculated Soil Organic Carbon Density dataset for China, addressing a specific need for standardized, depth-explicit, and temporally distinct density products. This focus on a robust SOCD calculation, combined with our rigorous methodology and specific spatio-temporal coverage, provides a unique and valuable resource for national-level carbon accounting, land management, and climate change research. We hope this clarification adequately explains the necessity and distinct contribution of our dataset. We've **already provided a clearer elucidation of this point in the manuscript**. If you still have questions, we would be very happy to offer a more detailed explanation.

4)Section 3.2: Have you taken into account the auto-correlation or interdependence among the potential predictors?

Thank you for your pertinent question regarding our consideration of autocorrelation and interdependence among potential predictor variables in our model. This is indeed a critical aspect of robust environmental modeling, and we have implemented a comprehensive and rigorous strategy to address both.

Our updated methodology incorporates a multi-stage feature selection approach specifically designed to optimize prediction accuracy while simultaneously managing multicollinearity (interdependence) and selecting the most informative features:

Addressing Interdependence (Multicollinearity):

Initial Correlation Screening: We first constructed a Pearson correlation matrix for our candidate features. To reduce redundancy and mitigate high linear interdependence (multicollinearity) among predictors, we systematically removed variables with very strong absolute Pearson correlation coefficients ($|r|>0.95$).

Exhaustive Combinatorial Optimization: Following this, for the most informative features, we conducted an exhaustive combinatorial testing of all possible feature subsets. This rigorous process allowed us to identify the optimal combination of features that maximized the coefficient of determination (R^2), ensuring the best predictive performance from a parsimonious set. This final step implicitly considers the combined effect and interdependence of features on model

accuracy, selecting a set that works optimally together.

Addressing Autocorrelation of Predictors and Model Robustness:

While our feature selection primarily targets interdependence, it's important to note that Random Forest models are inherently robust to multicollinearity and tolerate some degree of autocorrelation among predictor variables. They operate by recursively partitioning data based on individual features, making them less susceptible to issues that plague linear models when predictors are highly correlated.

More critically, our stratified spatial K-fold cross-validation strategy is explicitly designed to assess the model's performance on spatially independent data. By dividing the entire study area into K spatially independent sub-regions, we ensure that the model's performance is evaluated on areas it has "not seen" during training. This approach directly accounts for the potential spatial autocorrelation of predictor variables by testing the model's ability to generalize beyond local spatial patterns, thus mitigating the risk of overestimating performance due to spatial dependencies in the input data.

Through this comprehensive, multi-stage feature selection and model validation approach, we identified seven key features—mean annual temperature (Tem), elevation, normalized difference vegetation index (NDVI), clay content (Clay), simple ratio index (SR), bare soil index (BSI), and slope—which collectively provide the best predictive performance. This strategy effectively balances model complexity with predictive power, directly addresses predictor interdependence, and ensures the model's robustness against spatial autocorrelation for reliable large-scale SOCD mapping.

5)Section 3.3: You have already included temperature and precipitation as predictors of Random Forest Model. Why use climate zoning as well? It's unreasonable! Probably climate zoning promotes the training efficiency, but that's because the Random Forest Model you developed in this study is not efficient enough to fully learn the relationships. If you try deep learning models or AI foundation models, you may find out that the training effiency with and without climate zoning will be approximate.

Thank you for your constructive comment regarding the use of climate zones in our study, particularly your point that temperature and precipitation, used for defining climate zones, were also features in our initial Random Forest (RF) model. We appreciate you highlighting this potential redundancy.

We'd like to clarify our approach and the refinements made during our model development. Through an improved and rigorous feature selection method, we've refined our optimal feature set to comprise seven variables, consisting of mean annual temperature (Tem), elevation, NDVI, clay

content (Clay), simple ratio index (SR), bare soil index (BS1), and slope.

It's important to note that after this refinement, only mean annual temperature is now directly included as a predictor in our RF model. Precipitation is no longer a direct feature in this final set of predictors.

The climate zoning, as detailed in Section 3.3, serves a distinct and crucial purpose. Its primary role is to quantify the broad differences in temperature and precipitation across China and to improve the accuracy of SOCD estimation by developing zonal models. As referenced by Tang et al. (2018), SOCD exhibits significant variations across different climatic zones in China due to diverse environmental factors. 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. This strategy acts as a geographical stratification, enhancing the model's ability to account for macro-climatic differences and leading to more accurate predictions at a regional scale.

We believe this refined approach, where climate zoning functions as a beneficial stratification strategy rather than merely replicating direct predictors, strengthens our methodology.

Thank you for your insightful comments, especially regarding our use of climate zoning alongside temperature and precipitation as predictors in the Random Forest (RF) model, and your thoughts on model efficiency. We appreciate you raising these important considerations.

We want to first clarify an update to our methodology. While our initial model considered 12 environmental factors including temperature and precipitation, 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 (BS1), and slope. Therefore, in our final RF model, only mean annual temperature is directly included as a predictor, and precipitation is no longer a direct feature.

Regarding your point about the necessity of climate zoning and its relation to model efficiency, we acknowledge that advanced deep learning or AI foundation models might indeed learn complex relationships without explicit geographic stratification. However, for our current Random Forest modeling approach, climate zoning serves a strategic purpose beyond mere training efficiency.

As described in Section 3.3, climate zoning is implemented to quantify broad differences in temperature and precipitation across China and to improve the accuracy of SOCD estimation by developing zonal models. Given the vast and diverse environmental conditions across China, as highlighted by Tang et al. (2018), SOCD exhibits substantial variations linked to distinct climatic regimes. By classifying China into four subzones based on multi-annual average temperature and

precipitation thresholds, we are essentially performing a geographical stratification. Within each of these climatically distinct subzones, we then develop separate, localized SOCD estimation models.

This strategic division allows our models to capture regional nuances more effectively. The 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. Furthermore, this approach helps to enhance accuracy. By focusing the model training on more homogeneous environmental conditions within each zone, we significantly improve the predictive accuracy of SOCD estimates at a regional scale.

Thus, the climate zoning acts as a crucial stratification strategy that accounts for macro-climatic differences. While an 'ideal' model might learn these stratifications implicitly, for our Random Forest framework, this explicit zoning provides a robust way to manage the inherent heterogeneity of the study area and ensures more accurate and reliable SOCD predictions. We believe this approach is reasonable and justifiable for the scope and objectives of our study.

Comprehensively respond to Reviewer 2 and Reviewer 3.

Response to Anonymous Referee #2 and #3

We appreciate the reviewers' insightful comments regarding the inclusion of climate zoning in our Random Forest (RF) model, especially given that temperature and precipitation were initially considered as direct predictors. We understand the concern that this might introduce redundancy or indicate a limitation in the model's ability to capture complex relationships independently.

We would like to clarify our approach and the subsequent refinements made to our feature selection process. Initially, our model considered 12 environmental factors: Clay content, Sand content, Temperature, Precipitation, NDVI, Elevation, Slope, Aspect, Topographic Wetness Index (TWI), Canopy Nitrogen Discrimination (CND), Actual Humidity (AH), and China Land Cover Dataset (CLCD).

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 (BS1), and slope.

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

The climate zoning, as described in Section 3.3, serves a distinct purpose. It is implemented

to quantify the broad differences in temperature and precipitation across China and to improve the accuracy of Soil Organic Carbon Density (SOCD) estimation by developing zonal models. As referenced by Tang et al. (2018), significant variations in SOCD are observed across different climatic zones in China due to diverse and complex environmental factors.

Our climate zoning approach, based on multi-annual average temperature and precipitation thresholds (400 mm for MAP and 10°C for MAT), categorizes China into four subzones (humid, semi-humid, semi-arid, and arid). 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, rather than assuming a single, universally applicable relationship across the entire diverse landscape of China.

While we acknowledge the theoretical point that highly efficient deep learning or AI foundation models might inherently learn these relationships without explicit zoning, our current RF approach benefits significantly from this stratification. By segmenting the study area based on fundamental climatic drivers and then applying a robust RF model within each segment, we enhance the model's ability to capture regional nuances and ultimately improve the accuracy of SOCD estimation. The climate zoning, in this context, acts as a geographical stratification strategy that accounts for macro-climatic differences, enabling more accurate predictions at a regional scale.

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.

6)Many results, such as comparison results, are unconvincing, since significance test is lacking.

Thank you for your valuable comment regarding the convincingness of our comparison results due to a perceived lack of significance testing. We appreciate this point and would like to clarify our approach to statistical assessment.

We have indeed conducted robust statistical assessments to evaluate the significance of our findings. We utilized the **95% confidence intervals (CIs) of the Root Mean Square Error (RMSE)**. Crucially, these CIs were rigorously computed using the **Bootstrap Confidence Interval method**. The Bootstrap method is a powerful non-parametric resampling technique that allows for robust estimation of confidence intervals without making assumptions about the underlying data distribution, thereby enhancing the reliability of our statistical inferences.

As detailed in the manuscript (e.g., in Section 4.2, specifically when comparing the global model with the climate-zoned model), we explicitly demonstrate that:

- For the 0-20 cm depth, the RMSE 95% CIs of the global model ([2.13, 2.35] kg C/m²) and the climate-zoned model ([1.85, 1.99] kg C/m²) are non-overlapping. This clearly indicates a statistically significant improvement in RMSE achieved by implementing the climatic zoning strategy.
- Similarly, for the 0-100 cm depth, the RMSE 95% CIs of the global model ([7.34, 8.67] kg C/m²) and the climate-zoned model ([6.49, 7.78] kg C/m²) are also distinct, confirming the statistical significance of the performance enhancement from zoning.

The use of non-overlapping bootstrap-derived confidence intervals is a widely accepted and robust statistical method to infer significant differences between model performances. Our results, as demonstrated by these CIs, consistently show that our proposed climate-zoning model offers statistically significant improvements over the global model, and its performance across different climate zones is clearly delineated by these confidence intervals. We believe that these statistically significant improvements, substantiated by the robust Bootstrap-derived CIs, render our comparison results convincing and robust.

We hope this clarification fully addresses your concern.

7)Section 4.3: You developed the SOC map for China, but only did the independent validation in Heihe River Basin, which is quite small, and the climate condition there is quite different from many other parts of China (e.g., southern China).

Thank you for your comment regarding the scope of our independent validation in Section 4.3. We appreciate you raising this point, as it provides an opportunity to clarify our comprehensive validation strategy.

We would like to emphasize that our validation efforts were not limited to the Heihe River Basin. While the Heihe River Basin indeed served as one important independent validation area, our study employed a multi-pronged validation approach across various spatial and temporal dimensions to rigorously assess our model's performance and generalizability across China. This included:

Independent Test Set Validation (Spatial Generalization): This is our primary spatial validation. We employed a stratified spatial K-fold cross-validation strategy, dividing the entire study area (China) into *K* spatially independent and non-overlapping sub-regions. This ensured that a portion of the data (the test fold) was always from geographically unseen areas during model training. The results from this rigorous cross-validation (presented in Section 4.2, Table 3) reflect the model's overall generalization ability across the diverse conditions of China, far beyond just the Heihe River Basin. Clarification on Heihe River Basin & Other Independent Samples: For further assurance of spatial generalization, we specifically utilized additional independent sample points from various ecological regions. This included the Heihe River Basin (as you noted), along with other dedicated sample points from China's terrestrial ecosystems, grasslands, and desert ecosystems. These geographically diverse independent samples (as discussed in Section 4.3.1 and presented in Figure 9 & 10) provide further confirmation of the model's accuracy and universality across different depths (0-20cm and 0-100cm) and diverse ecological systems in China.

Comparison with Published SOCD Products (Spatial and Temporal Consistency): We conducted extensive comparisons of our estimated SOCD maps with several well-established, publicly available SOCD products, both global and regional. This included comparisons with SoilGrids250m, GSOCmap, and HWSD v2.0 (Figure 11) for spatial consistency, and crucially, a detailed comparison with the SOC Dynamics ML dataset in China (Li et al., 2022) (Figure 12) for consistency across the 1980s, 2000s, and 2010s. This multi-product comparison served as a critical external validation, confirming the consistency and accuracy of our estimates against recognized benchmarks.

Aggregated Results Comparison: We also presented aggregated results of our estimated SOCD compared with previous investigations for China (Figure 13), providing a broader context for the overall consistency of our SOCD estimates at national scales and different depths.

These diverse validation methods collectively ensure a robust and comprehensive assessment of our model's performance, covering both spatial generalization (including geographically distinct independent samples across various ecosystems) and temporal consistency across China. We have clearly presented these different validation strategies and their results in Sections 4.2 and 4.3 of the manuscript.

We will ensure that the descriptions in Section 4.3 are even more explicitly linked to this multi-faceted approach to prevent any further misunderstanding.

8)Section 4.4: The relatively high correlation with existing datasets cannot justify that your dataset is much better than existing ones. Moreover, some of these existing datasets you compared with are global-scale maps.

Thank you for your constructive comments regarding Section 4.4, particularly about the interpretation of correlations with existing datasets and the choice of comparison products. We appreciate this opportunity to clarify our intent and the unique value of our dataset.

You are entirely correct that a relatively high correlation with existing datasets alone does not automatically justify our dataset being "much better." We agree that our primary purpose in comparing with established products is not to claim overall superiority across all aspects, but rather to demonstrate the consistency, plausibility, and external validity of our newly developed SOCD product. Showing strong correlations indicates that our estimated SOCD maps align well with recognized global and regional benchmarks, providing confidence in their general patterns and magnitudes.

Furthermore, these comparisons serve to highlight the complementary value and specific advantages of our dataset, which differentiate it from existing products. Our dataset offers:

Precise SOCD (Soil Organic Carbon Density): Unlike some datasets that provide SOC content or stock, our study focuses on rigorously converting SOC content to SOCD using bulk density and coarse fragment data, offering a more standardized and direct metric for carbon accounting.

High Spatial Resolution and Depth Specificity for China: While some comparison datasets are global, our product provides 1-km resolution SOCD specifically for China, across specific depths (0-20 cm and 0-100 cm), which often represents a higher level of regional detail or a different depth range compared to the global products (e.g., GSOCmap only to 30 cm).

Multi-Decadal Long-Term Series: Our dataset provides a continuous time series of SOCD for China from 1985 to 2020 (in 5-year intervals), offering a unique temporal perspective that many existing static or shorter time-series products do not.

Robust Methodology: Our methodology, including the three-stage feature selection and stratified spatiotemporal cross-validation, contributes to the reliability and generalizability of our product.

Regarding your second point about comparing with global-scale maps, we acknowledge that their scope is different. However, comparing our regional product with global datasets like SoilGrids250m, GSOCmap, and HWSD v2.0 is highly valuable for external benchmarking. These global maps represent widely accepted standards in the field, and showing consistency with them helps to contextualize our regional product within a broader global framework. Crucially, we also included a direct comparison with the China-specific SOC Dynamics ML dataset (Li et al., 2022),

which provides a more direct regional benchmark for our methodology and temporal consistency.

In summary, the comparisons in Section 4.4 are designed to demonstrate our dataset's consistency, reliability, and unique contribution (in terms of precise metric, spatial/temporal resolution for China, and methodological rigor) rather than solely asserting its superiority over established global or regional products. We believe this comprehensive approach strengthens the overall justification for our dataset.

9)Section 4.5: Have you compared the temporal variation of SOC to other existing datasets or in-situ measurements? Can you validate your SOC temporal variation? Can you justify that your SOC dataset is much better than existing SOC maps in terms of temporal variation?

Thank you for your question regarding the validation of temporal variation in our SOC dataset in Section 4.5. We appreciate the opportunity to clarify this crucial aspect of our work.

We would like to confirm that we have indeed comprehensively validated the temporal variation of our estimated SOCD in Section 4.5, by comparing it with both existing datasets and a range of published investigations for China. Our approach is designed to provide robust evidence for the reliability of the temporal changes captured by our model.

Specifically, the temporal validation is demonstrated through:

Comparison with Aggregated Results from Previous Investigations (Figure 13): As presented in Figure 13, we directly compared our estimated SOCD changes over time (from the 1980s to the 2010s) with aggregated results from numerous published investigations in China (e.g., Ni, 2001; Wu et al., 2003; Wang et al., 2004; Xu et al., 2018; Wang et al., 2021; Li et al., 2022; Zhang et al., 2023). Figure 13 explicitly shows that our estimated SOCD values for both 0-20 cm (Fig. 13a) and 0-100 cm (Fig. 13b) fall within the ranges reported by these independent studies across different time points. This strong agreement, particularly the capture of the slight upward trend in 0-20 cm topsoil from the 1980s to the 2010s (consistent with soil management practices and environmental changes), directly validates the overall temporal dynamics and magnitudes of our estimated SOCD against a collective body of research.

Comparison with SOC Dynamics ML Dataset (Figure 12): Furthermore, as discussed in Section 4.3.2 and presented in Figure 12, we performed a direct visual and spatial comparison of our decadal SOCD maps (for 1980s, 2000s, and 2010s) with the China-specific SOC Dynamics ML dataset (Li et al., 2022). This inter-product comparison serves as a direct validation of our dataset's temporal patterns and magnitudes against another recognized regional product that explicitly models SOC dynamics. The consistency observed in this comparison strengthens the plausibility of the temporal changes derived from our model.

In addition to these direct comparisons, the temporal stratification embedded within our stratified spatial K-fold cross-validation (discussed in our response to RC2-4) inherently contributes to the validation of our model's ability to capture temporal variation. By ensuring that samples from all observed decades are proportionally represented in both training and validation sets, the model learns to generalize and predict across different historical contexts.

Regarding the justification of whether our dataset is "much better" in terms of temporal variation, we focus on highlighting its unique contributions and reliability rather than outright

superiority. Our dataset provides a comprehensive, high-resolution (1-km) and continuous multi-decadal time series (1985-2020 in 5-year intervals) of SOCD for China, generated by a single, unified space-time model designed to capture these dynamics. The robust consistency observed through comparison with existing investigations and the Li et al. (2022) dataset, combined with our rigorous methodology, confirms the reliability of the temporal changes in our product. We believe this makes our dataset a highly valuable and reliable resource for studying SOC dynamics in China over the long term.

1)RMSEs should have units

Thank you for your careful review and for pointing out the omission of units for RMSE values. We recognize the importance of precise reporting for all statistical metrics, and we appreciate you highlighting this detail, which enhances the clarity and interpretability of our results.

We have thoroughly reviewed the manuscript to address this oversight. We will ensure that the appropriate unit for Soil Organic Carbon Density (SOCD), which our RMSEs represent, is consistently and clearly added wherever RMSE values are reported. Our SOCD is quantified in kg C /m², and this unit will now accompany all RMSE values throughout the paper.

Specifically, these revisions will be implemented in the following sections:

The Abstract will be updated to include the unit if any RMSE value is mentioned there, ensuring our key performance indicators are immediately clear. In Section 4.2, 'Model Performance and Cross-Validation Results,' where we detail the predictive accuracy of our models, all RMSE values presented in the main text will explicitly state their unit. Furthermore, Table 3, which summarizes the cross-validation statistics, will have its column headers or relevant entries updated to clearly indicate that RMSE is reported in kg C /m². Moving to Section 4.3, 'Independent Validation Results,' any discussion of RMSE values, particularly those pertaining to independent test sets or comparisons with independent observed data (such as the Heihe River Basin data or Xu's measurements), will now include the kg C /m² unit. This will also apply to the captions or labels within any associated figures (e.g., Figures 9 and 10) that present RMSE or similar error metrics. Additionally, in Section 4.4, 'Comparison with Existing Datasets,' where we discuss the performance of our model relative to other published products, if RMSE is used as a comparative metric, its unit will be consistently provided. Lastly, should any RMSE values be referenced or re-discussed in the Discussion Section or other parts of the manuscript, we will ensure their units are correctly specified.

We are confident that these comprehensive adjustments will significantly improve the precision and readability of our results.

2)Lines 239-240: what does ‘before zone’, ‘after zone’ mean? Please polish your writing.

Thank you for your meticulous review and for highlighting the ambiguity in the phrasing "before zone" and "after zone" on lines 239-240. We sincerely apologize for this unclear expression and have undertaken a thorough optimization of our manuscript to ensure clarity and professionalism in all our descriptions.

Our original intention at this point was to illustrate the difference in model performance before and after the application of our climatic zoning strategy. Specifically, we have implemented the following explicit improvements in the manuscript:

Optimization of "before zone": We have consistently replaced the original "before zone" with clearer phrases such as "without climatic zoning" or, more directly, "when run as a global model." This clearly indicates the model's performance when trained uniformly across the entire Chinese region without considering climatic divisions.

Optimization of "after zone": Correspondingly, we have replaced "after zone" with "after implementing the climatic zoning strategy" or "based on the climatic zoning model." This explicitly refers to the improved performance observed after training separate sub-models for different climatic zones.

Taking your mentioned original sentence as an example: "The 0-100 cm SOCD prediction model has an accuracy of R2=0.44 and RMSE=8.09 before zones and R2=0.52 and RMSE=6.50 after zones, with R2 increased by 0.08 and RMSE decreased by 1.59."

We will revise it as follows: "The 0-100 cm SOCD prediction model achieved an accuracy of R^2=0.44 and RMSE=8.09 without climatic zoning (i.e., when run as a global model). This performance significantly improved to R^2=0.52 and RMSE=6.50 after implementing the climatic zoning strategy, resulting in an R^2 increase of 0.08 and an RMSE decrease of 1.59."

Through these revisions, readers will immediately understand the meaning of these two model states and their performance differences, thus better appreciating the advantages of our proposed climatic zoning algorithm. We have systematically checked and corrected all similar ambiguities throughout the manuscript, ensuring consistency and clarity in terminology usage.

3)Line 231: R2 for 0-20 cm is 0.43-0.59; R2 for 0-100 cm is 0.50-0.54. How can you conclude that the fitting or correlation is slightly worse for 0-100 cm compared to 0-20 cm?

Thank you for your very keen observation and rigorous questioning regarding the interpretation of data on Line 231. This point you've raised is crucial, as it prompts us to articulate the nuances of our model's performance with greater precision, thus preventing any potential misinterpretations.

You are correct that the R^2 ranges provided (0-20 cm: 0.43-0.59; 0-100 cm: 0.50-0.54) indeed show an overlap, and the minimum R^2 for 0-100 cm is even higher than that for 0-20 cm. We acknowledge that, solely based on these range values, readers might question our conclusion that "fitting or correlation is slightly worse for 0-100 cm."

However, our conclusion was not solely drawn from a literal comparison of these overall R^2 ranges. Instead, it is based on the final and best performance achieved by our optimized zoning model. As you may have noted in other sections of the manuscript (e.g., our summary of optimized model performance), after implementing the climatic zoning strategy, the peak R^2 for the 0-20 cm depth model reached 0.55, while for the 0-100 cm depth model, it was 0.52. Therefore, our statement of "slightly worse" refers to the ultimate model accuracy achieved through our best methodology, implying that the explanatory power of the 0-100 cm depth model, even at its best, was marginally lower than that of the 0-20 cm model. The R^2 ranges (e.g., 0.43-0.59) primarily reflect the variability or spread of model performance across different climatic zones or cross-validation folds, rather than the single point performance of the final selected model.

This subtle difference in performance also aligns with general understanding and scientific principles in soil organic carbon modeling:

Complexity and Accessibility of Driving Factors: Surface SOC (0-20 cm) dynamics are strongly influenced by factors like climate (temperature, precipitation), vegetation input, land use, and agricultural management practices. These factors are typically well-captured by surface-observable covariates derived from remote sensing and meteorological data. In contrast, deeper SOC (0-100 cm) is affected by more complex, long-term biogeochemical processes, 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 the types of macro-scale environmental covariates commonly employed in regional mapping studies.

Data Representativeness and Uncertainty: Measured data for deeper SOC are generally sparser and may have higher inherent variability compared to topsoil samples, which contributes to greater uncertainty in model predictions for these depths.

To convey this information more clearly, we will refine the phrasing on Line 231 and any related statements in the manuscript. We will explicitly state that the conclusion of "slightly

worse" is based on the comparison of the optimal performance of the zoning models and will briefly explain the inherent challenges in modeling deeper SOC. We are confident that these improvements will ensure readers accurately interpret our findings.

4)Lines 236-238: Suggest using relative RMSE (rRMSE, RMSE/mean value) instead of RMSE, since the SOC in arid regions can be quite low.

Thank you for your valuable suggestion regarding the use of relative RMSE (rRMSE) instead of RMSE, especially considering the potentially low SOC values in arid regions. We acknowledge that rRMSE can indeed offer valuable insights into model performance, particularly when dealing with variables that have varying magnitudes. **While rRMSE provides a useful normalized perspective, after thorough consideration and aligning with our specific research objectives and broader comparability needs, we have opted to primarily report RMSE for the following reasons:**

1. **Direct Measure of Absolute Error:** RMSE directly quantifies the absolute error between predicted and observed values. This provides a straightforward measure of the model's predictive accuracy in the original units (kg C/m²), which is crucial for understanding the practical significance of prediction errors in terms of carbon stock. While rRMSE normalizes this error, RMSE offers a clearer understanding of the actual magnitude of deviation.

2. **Extensive Comparability with Existing Research:** RMSE is a widely adopted and standard metric in numerous studies on soil organic carbon estimation and spatial modeling. Using RMSE facilitates direct comparison of our model's performance with a vast body of existing literature, thereby enhancing the broader applicability and contextualization of our findings within the scientific community. **Many authoritative studies similar to our research direction, such as those quantifying changes in soil organic carbon density using random forest models (Chen et al., 2023) and exploring the spatial patterns and controlling factors of soil organic carbon density (Huang et al., 2024), commonly employ RMSE as a key indicator. Furthermore, RMSE is a standard evaluation metric when comparing with existing SOCD products (Li et al., 2022; Xu et al., 2018; Dong et al., 2024).**

3. **Demonstrated Robust Performance:** As presented in our results, even in regions with potentially low SOC, our model exhibits consistently low RMSE values across different depths and climatic zones (e.g., our model's RMSEs for 0-20 cm and 0-100 cm in arid zones are 1.61 kg C/m² and 3.17 kg C/m², respectively). These results indicate a strong predictive capability, suggesting that RMSE adequately captures the model's accuracy and precision across the study area, including arid regions. The low absolute errors, as reflected by RMSE, demonstrate the effectiveness of our approach.

Therefore, we believe that RMSE serves as a robust and appropriate metric for evaluating our model's performance, effectively reflecting its absolute predictive accuracy and facilitating meaningful comparisons with previous studies.

5)Lines 242-244: I thought the performances are similar. Have you performed significance test? In addition, the explanation is not quite convincing. Can you provide more robust proof?

Thank you for your insightful comment regarding the perceived similarity in model performances and the need for significance testing and more robust proof. We appreciate you raising these points, and we have taken significant steps to address them in the revised manuscript.

We understand that the original phrasing in lines 242-244 might have led to the perception of similar performances without sufficient statistical backing. To clarify this, we have thoroughly **revised the explanation** in the corresponding sections (e.g., Section 4.2 and the detailed discussion of climate zone-specific performance in Section 4.2.1, previously supplied). More importantly, we have **explicitly incorporated statistical significance testing** through the rigorous calculation of **95% Confidence Intervals (CIs) for RMSE, derived using the Bootstrap method**.

The Bootstrap method is a powerful non-parametric technique that allows for robust estimation of CIs without assuming data distribution, providing a strong basis for statistical inference. We use the non-overlapping nature of these CIs to infer statistically significant differences between model performances:

- When the 95% CIs of two performances (e.g., from different models or climate zones) **do not overlap**, it indicates a statistically significant difference.
- Conversely, when the 95% CIs **overlap**, it suggests that any observed differences are not statistically significant at the 95% confidence level, thus statistically confirming a "similar" performance.

As now detailed in the revised manuscript, particularly in the discussion of model performance across different climate zones (corresponding to the content previously provided for Figure 7):

- For example, at the 0-20 cm depth, while the humid and semi-arid zones show very close RMSE values (1.77 kg C/m² for both), their respective RMSE 95% CIs ([1.61, 1.93] and [1.49, 1.63]) allow for a robust assessment of their similarity. In contrast, the RMSE CI for the arid zone ([1.17, 2.07]) is distinct from that of the semi-humid zone ([2.39, 2.76]), demonstrating a statistically significant difference.
- Similarly, for the 0-100 cm depth, we demonstrate significant differences in performance, such as the arid zone exhibiting the lowest RMSE with a tight CI ([2.56, 3.81]), confirming good precision despite a lower R^2.

Regarding the explanation's convincingness, the **revised discussion** (which now includes details on data distribution, environmental consistency, and unique hydrological/biological factors

in drylands) is now directly supported by these **quantitative statistical proofs (CIs)**. The CIs provide objective evidence for why performances are distinct or similar, lending much stronger support to our interpretations of the underlying environmental controls.

We believe that the combined efforts of refining the descriptive explanations and introducing the Bootstrap-derived 95% CIs for robust statistical validation have significantly enhanced the clarity, rigor, and convincingness of our comparison results.

We hope this detailed response fully addresses your valuable concerns.