Response to Reviewers' comments to manuscript ESSD-2025-264

"Spatial Patterns of Sandy Beaches in China and Risk Analysis of Human Infrastructure Squeeze Based on Multi-Source Data and Ensemble Learning"

Dear Reviewers:

Thank you very much for your thoughtful and detailed review. Your suggestions have provided us with important and constructive insights, which have significantly improved the manuscript. We have carefully considered all of your comments and have made substantial revisions to the manuscript based on your feedback. We have done our best to enhance the manuscript and hope that the revised version will meet your approval. A point-by-point response to the outstanding comments is attached to this manuscript. The major revisions are summarized as follows:

Response to Comments by Reviewer 1:

1. *After carefully reading the revised manuscript and the authors' responses, I recommend the paper for publication, provided that a few additional minor revisions are addressed. The authors have satisfactorily responded to most of the comments received, leading to an overall improvement of the manuscript, particularly through the use of more appropriate metrics to assess model performance. However, I have several remaining general comments that should be considered before final acceptance. I also attach an annotated version of the manuscript with further minor remarks and typographical corrections that should assist the authors in preparing the final version.*

Response:

We would like to express our gratitude to the reviewer for their valuable comments. Based on the suggested revisions in the attached document, we have made the following adjustments to the manuscript:

1. We have added "**(VNTL)**" at **L120**, corresponding to VNTL in the NDUI formula in **Table 2**. The nighttime light data are used in the calculation of NDUI, which can be employed to distinguish coastal buildings from sandy beaches.

2. L17: We have changed "**single-year**" to "**unique**" in the manuscript.

3. L19: We have modified the sentence from: "**(2) In Fujian, Guangdong, and Taiwan, the identified sandy beaches covered 180.05 km², with perimeters of 5610.26 km and widths of 54.91 m, 38.92 m, and 57.17 m, respectively.**" to: "**(2) In Fujian, Guangdong, and Taiwan, the identified sandy beaches covered 54.57 km², 78.88 km², and 46.60 km², with perimeters of 1435.89 km, 2849.39 km, and 1324.98 km, and widths of 54.91 m, 38.92 m, and 57.17 m, respectively.**" in the manuscript.

4. L42: We have changed "**unmanned aerial vehicles (UAVs), or aerial platforms**" to "**and aerial platforms—including unmanned aerial vehicles (UAVs)**" in the manuscript.

5. L62: We have modified the sentence from: "**For example, Latella et al. conducted a exploratory survey and monitoring of sandy beaches by comparing Sentinel-2 and Landsat images, using random forests and various spectral indices (Latella et al., 2021); while Yong et al. used a binary image segmentation method based on the U-Net model in convolutional neural networks to accurately delineate the sandy beach outline of the southeastern coast of Australia (Yong et al., 2024).**" to: "**For example, an exploratory survey and monitoring of sandy beaches was carried out by comparing Sentinel-2 and Landsat images, using random forests and various spectral indices (Latella et al., 2021), while a binary image segmentation method based on the U-Net model in convolutional neural networks was applied to accurately delineate the sandy beach outline of the southeastern coast of Australia (Yong et al., 2024).**" in the manuscript.

6. L86: We removed "**build**" and retained "**construct**" in the manuscript.

7. L135: We have modified the sentence from: "**(2) The 2022 China 10m sandy beach dataset identified by Ni et al. (Ni et al, 2024) using a support vector machine method based on Sentinel-2 imagery; (3) The 2020 China coastal land use dataset at 10m resolution, identified by Miao et al. (Miao et al, 2022) using an object-oriented classification method based on Sentinel-2 imagery.**" to: "**(2) The 2022 China 10 m sandy beach dataset was identified using a support vector machine method based on Sentinel-2 imagery (Ni et al, 2024); (3) The 2020 China coastal land use dataset at 10 m resolution was identified using an object-oriented classification method based on Sentinel-2 imagery (Miao et al, 2022).**" in the manuscript.

8. We added a detailed explanation of the division of samples into training and testing sets in **L157**. The original sentence reads: "**The samples were randomly divided into training and testing sets at an approximate 7:3 ratio, while ensuring consistent sample proportions across the 12 regions of the study area and maintaining balanced ratios of sandy beach and non-sandy beach samples within each region. This partitioning strategy effectively avoided regional bias and ensured the reliability and representativeness of the model evaluation.**" Additionally, the caption of **Fig. 4** was changed from "**Training and testing sets.**" to "**Annual counts of sandy beach and non-sandy beach samples in the training and testing sets.**" in the manuscript.

9. L171: We have changed all the fonts in **Table 2** to **Times New Roman**.

10. We have added detailed explanations of the five evaluation metrics (Accuracy, Precision, Recall, sandy beach F1-score, and Area Under the Curve (AUC)) in **L193** and **L200**. The added content is as follows: "**These metrics were chosen to comprehensively evaluate different types of classification errors and overall model performance. Accuracy measures the proportion of correctly classified samples, reflecting overall correctness. Precision indicates the reliability of**

positive predictions, highlighting false positives, while Recall reflects the model's ability to detect all positive samples, highlighting false negatives. The F1-score balances Precision and Recall, providing a single metric when both false positives and false negatives are important. AUC evaluates the model's discrimination ability across different thresholds, indicating stability and robustness." and "**TP (True Positive) refers to the number of samples that are actually positive and correctly classified by the model, while TN (True Negative) refers to the number of samples that are actually negative and correctly classified. FP (False Positive) represents the number of samples that are actually negative but incorrectly classified as positive, and FN (False Negative) represents the number of samples that are actually positive but incorrectly classified as negative. TPR (True Positive Rate, or Recall) measures the model's ability to identify positive samples and is calculated as** $TPR=\frac{TP}{TP+FN}$, **whereas FPR (False Positive Rate) measures the proportion of negative samples incorrectly classified as positive, calculated as** $TPR=\frac{FP}{FP+TN}$."

11. L207: We have added the content on importance analysis. The newly added text is as follows: "**SHAP (Shapley Additive exPlanations) values are a model interpretation method based on cooperative game theory, used to quantify the contribution of each feature to the model's prediction. They calculate the marginal contribution of each feature across all possible feature combinations and take the average, thereby fairly attributing the "responsibility" for the prediction. This approach can explain the prediction behavior for individual samples or the entire model. In this study, SHAP values were used to interpret the contributions of a stacked ensemble model. Specifically, SHAP values were first calculated for each base learner (RF, GBDT, XGB, LGBM), then these values were normalized and weighted according to the coefficients of the meta-learner (LR) to obtain the contribution of each base learner to the final ensemble prediction. Finally, the weighted SHAP values of all base learners were summed to derive the SHAP values for the entire stacked ensemble model, thereby quantifying the influence of each base model on the final prediction and providing a measure of interpretability for the ensemble learning.**"

12. L217: We have modified the sentence from: "**Approximately 33% of global sandy beaches lack more than 100 meters of infrastructure-free space, and in this study, human infrastructure squeeze risk is defined as the risk that occurs when infrastructure (such as buildings, roads, ports, etc.) is located within 100 meters of a sandy beach area.**" to: "**Approximately 33% of global sandy beaches have less than 100 meters of infrastructure-free buffer space, and in this study, human infrastructure squeeze risk is defined as the risk that occurs when human infrastructure (such as buildings, roads, ports, etc.) is located within 100 meters of a sandy beach area**." in the manuscript.

13.L234: We have changed "**merged data**" to "**maximum-extent composite**" in the manuscript.

14.L241: We have changed "**has the most**" to "**hosts the largest number of**" in the manuscript.

15. Regarding Figure 6: The current binning scheme already provides a reasonable and sufficient representation of the statistical distribution. Further refinement of the bins would neither alter the overall pattern nor provide additional meaningful insights. In addition, following the reviewer's

suggestion, we have supplemented the figure caption by adding the following sentence at **Line 251**: "**The histograms in panels (b), (c), and (d) are based on the results of each individual beach rather than regional statistics.**"

16. L253: We have changed "**the variable importance of the model**" to "**the importance of model variables**" in the manuscript.

17. L264: We have changed "**SVM**" to "**RF**" in the manuscript.

18. We assessed this based on a comprehensive comparison with **Figs. 8 and 9**. **Fig. 8** provides Sentinel-2 true color imagery, and we judged the closeness to actual values by comparing our identified sandy beaches with the corresponding features in the true color imagery. In addition, we have added the phrase "**as shown in Figs. 8 and 9**," at **Lines 281 and 286** to clarify this in the manuscript.

19. L285: We added a "**than**".

20. L286: We have changed "**datasets**" to "**dataset**" in the manuscript.

21. Regarding Figure 10: We added a legend to the figure.

22. L319: We added "**, taking 2024 as an example.**"

23. L401: We have changed "**beach extraction**" to "**sandy beach extraction**" in the manuscript.

24. L420: We have changed "**sandy beach regions**" to "**sandy beaches**" in the manuscript.

2. *First, most figure captions could be strengthened, as several contain repetitions or lack essential detail, making them overly simple or general. The same applies to the table titles. Additionally, not all parameters and variables are introduced in the text (e.g., those appearing in Tables 2–4), and these should be clearly defined.*

Response:

Thank you for your valuable feedback; it has been extremely helpful. I will now respond to each of your comments:

We appreciate your suggestion. We believe that Tables 2 and 3 are already described in sufficient detail. Many variables in Table 2 are standard metrics commonly used in the field, and defining each individually in the text would substantially increase the manuscript length without adding significant value. Similarly, Table 3 does not include detailed explanations of the base learners because these algorithms are widely used and well-known; providing full definitions would take up considerable space without meaningful benefit.

Nevertheless, we have provided detailed explanations for Table 4, as indicated in Lines 193 and 200, to clarify the model evaluation metrics and their calculation.

At L193:

"**These metrics were chosen to comprehensively evaluate different types of classification errors and overall model performance. Accuracy measures the proportion of correctly classified samples, reflecting overall correctness. Precision indicates the reliability of positive predictions, highlighting false positives, while Recall reflects the model's ability to detect all positive samples, highlighting false negatives. The F1-score balances Precision and Recall, providing a single metric when both false positives and false negatives are important. AUC evaluates the model's discrimination ability across different thresholds, indicating stability and robustness.**"

At L200:

"**TP (True Positive) refers to the number of samples that are actually positive and correctly classified by the model, while TN (True Negative) refers to the number of samples that are actually negative and correctly classified. FP (False Positive) represents the number of samples that are actually negative but incorrectly classified as positive, and FN (False Negative) represents the number of samples that are actually positive but incorrectly classified as negative. TPR (True Positive Rate, or Recall) measures the model's ability to identify positive samples and is calculated as** $\mathbf{TPR=\frac{TP}{TP+FN}}$**, whereas FPR (False Positive Rate) measures the proportion of negative samples incorrectly classified as positive, calculated as** $\mathbf{TPR=\frac{FP}{FP+TN}}$**.**"

3. *As noted in the annotated manuscript, the authors mention merging annual datasets to create a single dataset, but the procedure used is not described. This is important methodological information and must be included. For example, were only sandy beaches identified in all years retained?*

Response:

Thank you for your valuable feedback; it has been extremely helpful. I will now respond to each of your comments:

The merging procedure is straightforward, using a **maximum-extent composite**. This has been clarified in **Line 234**, with the revised sentence as follows: "**Based on the maximum-extent composite of sandy beach recognition results over 8 years, the spatial distribution of the number, length, width, and area of sandy beaches in China is shown in Fig. 5.**"

4. *Regarding my earlier comment in the first review round, the authors have chosen different or additional metrics to evaluate model performance. However, the manuscript would benefit from a brief explanation of these metrics: why they were selected, their relevance to the study, and whether they provide complementary insights.*

Response:

Thank you for your valuable feedback; it has been extremely helpful. I will now respond to each of your comments:

We have provided detailed explanations for Table 4. Specifically, the following descriptions have been added at L193 and L200, respectively:

At L193:

"**These metrics were chosen to comprehensively evaluate different types of classification errors and overall model performance. Accuracy measures the proportion of correctly classified samples, reflecting overall correctness. Precision indicates the reliability of positive predictions, highlighting false positives, while Recall reflects the model's ability to detect all positive samples, highlighting false negatives. The F1-score balances Precision and Recall, providing a single metric when both false positives and false negatives are important. AUC evaluates the model's discrimination ability across different thresholds, indicating stability and robustness.**"

At L200:

"**TP (True Positive) refers to the number of samples that are actually positive and correctly classified by the model, while TN (True Negative) refers to the number of samples that are actually negative and correctly classified. FP (False Positive) represents the number of samples that are actually negative but incorrectly classified as positive, and FN (False Negative) represents the number of samples that are actually positive but incorrectly classified as negative. TPR (True Positive Rate, or Recall) measures the model's ability to identify positive samples and is calculated as** $\mathbf{TPR=\frac{TP}{TP+FN}}$**, whereas FPR (False Positive Rate) measures the proportion of negative samples incorrectly classified as positive, calculated as** $\mathbf{TPR=\frac{FP}{FP+TN}}$**.**"

5. *Finally, the manuscript reports different quality scores for the models, raising questions about whether the same evaluation procedure was used throughout. For example, the quality scores presented in Table 5 differ from those shown in Table 7 (S + T + Tr + P) and Figure 13 (Stacking).*

If the same model and inputs were used, these scores would be expected to align. The authors should clarify why the quality scores differ across these results.

Response:

Thank you for your valuable feedback; it has been extremely helpful. I will now respond to each of your comments:

We appreciate the reviewer's comment. It appears there is a misunderstanding. In the manuscript, the feature combination used is (**S + I + T + Tr + P**). Additionally, the results presented in **Table 7** and **Figure 13** are illustrative examples for the year **2024**, showing the model performance specifically for that year. These results are consistent with the **2024** results reported in **Table 5**. Therefore, the quality scores are indeed aligned when considering the same year and feature combination.