Responses to RC #2

Dear Reviewer,

Thank you very much for your helpful comments. The manuscript has been fully revised according to your suggestions. The following is a point-to-point response to your comments, and responses are in **blue.**

General comments：

The manuscript presents the development of a global plastic greenhouse distribution map (Global-PCG-10) based on Sentinel-2 10 m imagery. Weak labels were generated via a random forest classifier, and these were refined through a deep learning framework integrated with an active learning strategy. The spatial patterns, extent, and proportional coverage of global PCGs were analyzed, yielding results of clear scientific merit and practical utility. However, several issues and deficiencies must be addressed prior to publication:

Response:

Thank you for your encouragement and support on our study. We have revised the manuscript carefully according to your suggestions. The following is a point-to-point response to each suggestion.

1. Some in-text citations do not conform to the journal's formatting requirements. For example, in line 64 the citation "Aguilar et al., (2016)" incorrectly includes a comma. Please review and standardize all reference notations.

Response:

We appreciate your careful attention on the manuscript. As suggested, we have thoroughly checked all in-text citations and have revised to ensure consistency with the journal's formatting requirements. Specifically, we have corrected instances such as "Aguilar et al., (2016)" to "Aguilar et al. (2016)" and made similar adjustments wherever necessary.

2. In Figure 2 (Phase 2) the train/validation split is stated as 7:3, whereas in the main text it is described as 8:2. Please verify and ensure consistency in both places.

Response:

Thank you very much for pointing out this inconsistency. We feel sorry for our careless mistake. Actually, there is a mismatch between the figure and the description in the manuscript. We have revised the figure to ensure that the train/validation ratio is consistent with what is reported in the manuscript (8:2). We sincerely appreciate your careful review and helpful comment, which helped us improve the clarity and accuracy of our work.

3. In lines 162-163 authors state that the initial labels were generated via GEE and RF classification. Was any quality control applied to these labels (e.g., manual verification rate, error-rate assessment)? Please clarify.

Response:

We appreciate your concern regarding the quality control of the initial labels. Actually, prior to training the deep learning model, we implemented multiple strategies to ensure the accuracy and reliability of initial PCG labels, as detailed below.

(1) Collection of high-confidence samples.

To construct the training samples required for GEE-based Random Forest (RF) classification, we conducted field surveys in key greenhouse-intensive regions in China (e.g., Weifang in Shandong, Kunming in Yunnan, and Lishu in Jilin). During the surveys, we also consulted local farmers to confirm the locations and types of PCG. Considering that PCG typically remain in use for around 10 years or more with relatively high stability (Ou et al., 2021), we performed systematic manual visual interpretation of historical high-resolution imagery from Google Earth in multiple global regions to obtain high-confidence samples. For areas outside China, we additionally referred to published literature, meta-analyses and online sources for auxiliary identification. All samples were further verified using Sentinel-2 imagery to ensure their actual presence in the year 2020. We also refined PCG and non-PCG labels based on the RF classification results within each grid to enhance overall labeling accuracy.

(2) Quality assessment and selection of RF classification results.

Based on the collected samples, we trained a RF model with GEE, using a split between training and validation sets. A confusion matrix was constructed to evaluate the classification accuracy, where the validation set was excluded from the training process and used solely for accuracy assessment. Only those classification maps with an overall accuracy (OA) greater than 95% and a user accuracy (UA) for the PCG class above 90% were selected as candidate label maps for training the deep learning model.

(3) Final screening of training labels for the deep learning model.

All candidate label maps were undergone further visual inspection. Each 512×512 pixels image patch was visually checked to ensure high annotation quality, and only the regions with the most reliable classification results were retained for deep learning model training. This process ensured that the final training labels used in the deep learning model were of high reliability.

The above contents have been added in Section 3.1 “Stage-1: PCG weak label generation” of the manuscript (**see *Lines 178 ~ 200* for details**).

4. In lines 165-166 (Stage 3) mention “post-processing was applied to the PCGs classification results to

eliminate isolated noises" without specifying the algorithm or parameter settings. Please provide details of the post-processing method so that the results are reproducible.

Response:

We sincerely thank you for the valuable suggestion. In this study, we adopted a post-processing step that used a pixel-connected component-based Sieve Filter to eliminate isolated noises in the initial classification results. Specifically, we used the gdal.SieveFilter() function from the GDAL library (invoked in the Python environment) to perform the filtering. An 8-connected neighborhood was adopted, and a set of hierarchical thresholds for the minimum number of connected pixels (10 / 20 / 50) was applied. This multi-level threshold setting was designed to accommodate variations in noise distribution and mapping requirements across different regions.

The above contents have been added in Section 4.2 "Reliability of Global-PCG-10" of the manuscript (**see *Lines 455 ~ 458* for details**).

5. The manuscript alternates between "PCG" and "PCGs" (e.g., line 166 "PCGs classification results" vs. Figure 2 "PCG classification results"). Please unify the terminology throughout.

Response:

Thank you for your valuable comments and suggestions. We have carefully reviewed and revised the usage of "PCG" and "PCGs" throughout the manuscript. The terminology has been standardized as PCG accordingly.

6. In lines 239-243 outline the overall network architecture but omit implementation details of the MDCN and non-local modules. Schematic diagrams or detailed descriptions of these submodules are required.

Response:

We appreciate your valuable suggestion. In response, we have added detailed descriptions of both the MDCN and non-local modules in the revised manuscript. Additionally, we have provided schematic diagrams illustrating the structures of these submodules to improve clarity (see new Figure 5). We believe these additions could enhance the readability and completeness of the network architecture description.

The detailed reply is as follows.

In this study, we employed a deep semantic segmentation model, APC-Net (Niu et al., 2023a), as the core model to generate the final PCG classification map in a coarse-to-fine manner. APC-Net effectively integrates local and global features through multi-scale feature learning, thereby enhancing its classification capability under complex global terrain conditions.

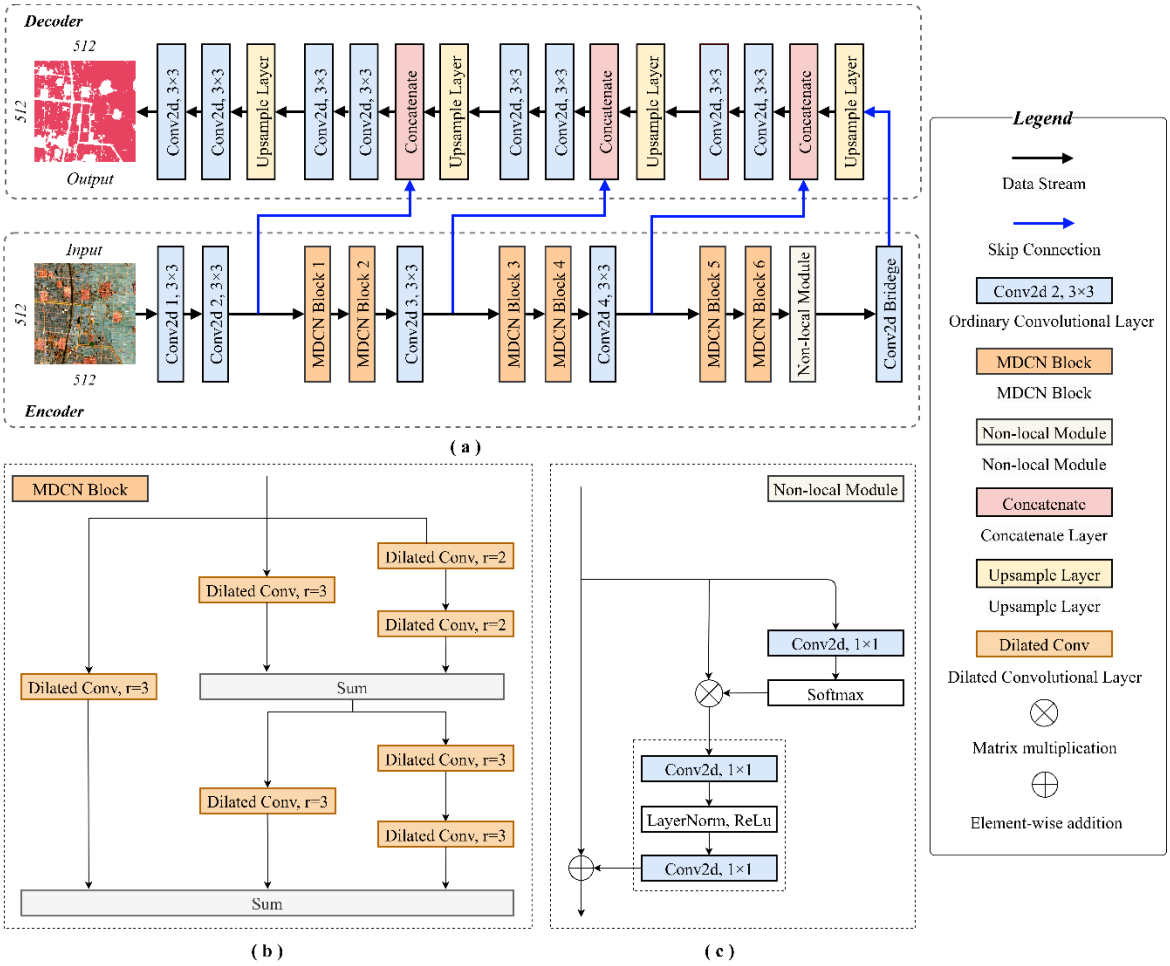

Figure 5. Overview of the proposed APC-Net model

Specifically, APC-Net consists of two main components, an encoder and a decoder (see Figure 5a). The encoder, which is the core of the network, takes a 512×512 remote sensing image patch as input and extracts highly representative features through a multi-layer structure. This not only enhances intra-class consistency but also improves inter-class separability. The encoder includes convolutional layers, an MDCN (Multi-scale Dilated Convolutional Network) module and a non-local module. The MDCN module (Figure 5b) integrates multi-scale dilated convolutions to effectively capture multi-scale local features, addressing scale variation issues that are common in PCG classification. The non-local module (Figure 5c) focuses on capturing global contextual information, thereby improving the model's capability for overall scene understanding. The decoder is responsible for restoring spatial information from the downsampled feature maps generated by the encoder and producing the final segmentation map. It employs bilinear interpolation for upsampling and skip connections to fuse encoder features, further refining the representation. The final PCG classification output maintains the same spatial resolution (512×512) as the input image.

The above contents have been added in Section 3.1.2 "APC-Net model" of the manuscript (**see *Lines***

***255 ~ 271* for details**).

7. Section 3.2.2 provides an overview of the active-learning strategy but lacks specifics on the number of iterations, per-iteration sample-selection criteria, and stopping conditions. Please supply quantitative stopping rules rather than the current qualitative phrase "until performance stabilizes or results are satisfactory."

Response:

Thanks for your valuable comment. In this study, the active learning process was conducted for up to five iterations. Each iteration involved a complete model training procedure, followed by performance evaluation using a validation set, focusing primarily on overall accuracy (OA) and mean Intersection over Union (mIoU). To determine whether to proceed with additional iterations, we applied a quantitative stopping criterion as follows. If the improvement in both OA or mIoU between two consecutive iterations was less than 1%, the model was considered to have reached performance stability, and the iterative process was terminated.

Regarding the sample selection strategy for each iteration, we identified candidate samples for manual review by comparing the model predictions with the initial weak labels. Specifically, an image patch was selected for further review if the Patch-level IoU between the model prediction and the weak label was less than 0.6. This threshold was determined empirically based on our observation of preliminary results and manual verification, aiming to balance classification accuracy and the efficiency of the active learning process.

The Patch-level IoU served as a metric to quantify the degree of disagreement between the model and the weak labels within a single patch, helping to identify potential labeling errors or uncertain regions. This patch-wise screening strategy enabled each iteration to focus on the most informative samples, thereby effectively improving both training data quality and model performance.

The above contents have been added in Section 3.2.2 "Active Learning strategy" of the manuscript (**see *Lines 284 ~ 300* for details**).

8. In lines 253-254, authors state that "if results are unsatisfactory, significantly altered initial labels are selected for further training," but "expected standards" and "significant alteration" are not quantified. Please define the thresholds and the human-involvement workflow in detail.

Response:

Thank you for your valuable comments. In response to your concern regarding the quantification of "expected standards" and "significant alteration," we have provided further clarification as follows.

(1) Expected Standards. In practice, we did not predefine rigid accuracy thresholds as expected

standards during model training. Instead, we primarily relied on visual interpretation of the PCG classification results to evaluate model performance. This approach is grounded in the following considerations. First, an early stopping mechanism was applied during deep learning model training to prevent overfitting. Second, the training and validation sets were split in an 8:2 ratio to maximize the use of sample data and improve model performance. Third, the training process involved iterative optimization of initial labels, allowing the model to learn more representative features and thus achieve improved PCG classification results.

Since it is difficult to predict the final performance of the model prior to training, we evaluated the results based on empirical experience, classification accuracy metrics and spatial distribution consistency. According to previous experience, an Overall Accuracy (OA) above 90% and a mean Intersection over Union (mIoU) above 0.6 are generally considered acceptable for binary classification tasks and served as our practical reference thresholds.

(2) Significant Alteration. "Significant alteration" refers to cases where there is a clear discrepancy between the initial labels and the model's predictions. Specifically, we quantified this by defining an image patch as significantly altered if the Patch-level IoU between the initial labels and model predictions is below 0.6. In such cases, the image patch is flagged for further review and potential label updating.

(3) Human-involvement Workflow. Once the model identifies image patches requiring label updates, the following process is initiated. The model calculates the Patch-level IoU between its predictions and the initial labels for each image patch. If the IoU is below 0.6, the image patch is flagged for manual review. Human annotators examine these image patches to verify whether the predicted labels align with the actual conditions. If the labels are found to be incorrect, they are manually corrected and returned for model retraining. The updated labels are then incorporated into the training dataset for the next iteration. This iterative process continues until either (a) the model reaches the expected performance standards, or (b) the improvement in either OA or mIoU across two consecutive iterations is less than 1%, the training process is terminated.

The above contents have been added in Section 4.2 "Reliability of Global-PCG-10" of the manuscript (**see *Lines 284 ~ 300* for details**).

9. In lines 270–275, authors note the use of the Adam optimizer and data partitioning but omit critical hyperparameters such as total epochs, batch size, and learning-rate decay strategy. Please provide these training details.

Response:

Thank you for the helpful comment. We have provided the missing training details, including the total number of epochs, batch size, learning rate, learning-rate decay strategy and early stopping criteria

in the revised manuscript. Specifically, the model was trained for 200 epochs with a batch size of 8, using an initial learning rate of 1e-4 with a step decay schedule. In addition, early stopping was applied with a patience of 10 epochs based on the validation loss to prevent overfitting. We believe these additions could improve the transparency and reproducibility of our work.

All of the above contents have been added in Section 3.2.3 "Training Details" of the manuscript (**see *Lines 314 ~ 318* for details**).

10. In lines 272-273, the dataset is described as comprising 14 825 training images and 3 707 validation images. Please explain how the sample was divided (randomly or by region? Is global diversity guaranteed?).

Response:

Thank you for your valuable comment. Considering that PCG areas in China account for more than two-thirds of the global total, an overrepresentation of Chinese samples in the training dataset may reduce the model's ability to accurately identify PCG in other regions. To mitigate this issue, we divided the samples into two subsets, one for China and the other for non-China regions, allowing the model to learn region-specific PCG features separately. To further enhance the global representativeness of the dataset, we included PCG distribution areas across multiple continents, including Asia, Europe, North America, South America, Africa and Oceania during the sample construction stage. This could ensure sufficient geographic diversity in both the training and validation sets, thereby improving the model's generalization capability.

Specifically, we collected 10,230 samples from China and 8,302 samples from other regions. Each subset was then randomly divided into training and validation sets using an 8:2 split to train two separate regional PCG classification models. The predictions of the two models were finally combined to generate the Global-PCG-10 dataset, a global 10-meter resolution PCG classification product.

All of the above contents have been added in Section 3.2.3 "Training Details" of the manuscript (**see *Lines 319 ~ 327* for details**).

11. In line 292, the term "false positives (FP)" is incorrectly written as "false negatives." This typo may confuse the metric definitions. Please verify that all terminology and formulas in the text and tables are consistent.

Response:

We appreciate your careful reading and helpful comment. We have corrected the typo at line 292, changing "false negatives" to "false positives (FP)" as intended. We have also carefully reviewed the entire manuscript, including all metric definitions, formulas and tables to ensure consistency and accuracy.

12. Section 4.2 presents only examples with no significant false alarms, which seems inconsistent with the overall recall rate of 84%. Please include error - case analysis and discuss the causes of missed detections.

Response:

We sincerely thank you for the constructive suggestion to include an analysis of missed detection (omission) cases. In response, we have added a detailed discussion of typical omission scenarios. The relevant description is as follows.

Based on the test samples and a systematic comparison with the 3-meter resolution PCG data provided by Tong et al. (2024), we identified two main types of omission errors in the current Global-PCG-10 dataset during the PCG extraction process, as detailed below.

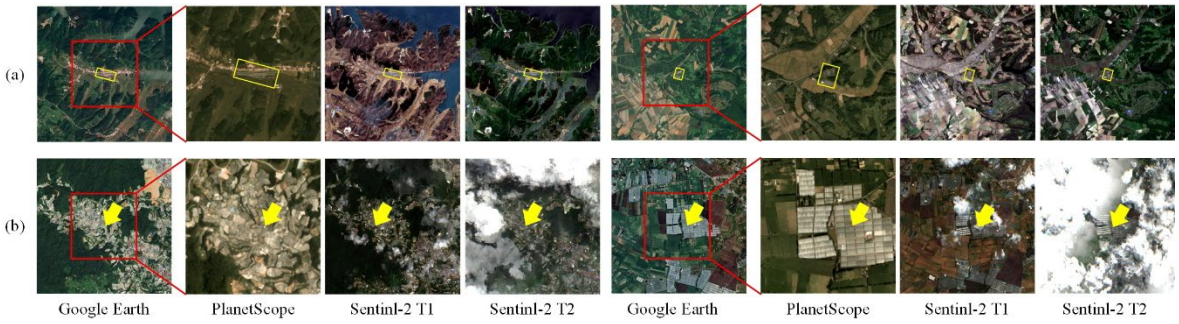

Figure 11. Bad case analysis

(Note: From left to right, the image shows the Google Earth imagery from 2020, 3-meter spatial resolution PlanetScope imagery from 2019 and Sentinel-2 imagery from spring and summer, respectively.)

As shown in Figure 11a and b, due to the relatively coarse spatial resolution (10 meters) of Sentinel-2 imagery compared to higher resolution sources such as PlanetScope or Google Earth (3 meters or finer), small-scale PCG targets often occupy only a few to a dozen pixels. These pixels are usually mixed pixels that contain spectral information from multiple surrounding land cover types. As a result, the model finds it difficult to extract PCG's distinct spectral and texture features, which impairs its ability to accurately detect small and visually inconspicuous PCG. For instance, in the area shown in Figure 11a, the PCG can be roughly identified in the high-resolution image, with some observable texture patterns. However, in the corresponding Sentinel-2 image at 10-meter resolution, the PCG contours are blurred and lack clear geometric and textural features, leading to missed detections.

Meanwhile, the Global-PCG-10 dataset is derived using multi-temporal Sentinel-2 imagery from spring and summer, organized by 1° grid tiles. However, due to cloud contamination and limited observation opportunities, it is challenging to obtain cloud-free images for both seasons in some regions (Figure 11b). This could limit the model's ability to extract consistent temporal features, thereby increasing the likelihood of omission errors. Figure 11b presents a typical case, although the overall cloud coverage is relatively low, even thin clouds can affect surface reflectance values and interfere with the

model's classification performance.

In summary, misclassification errors in PCG classification primarily arise from two aspects: (1) the presence of mixed pixels in medium-resolution imagery when detecting small-scale PCG, which weakens the model's ability to learn effective spectral and textural representations; and (2) limitations in the spatial and temporal availability of remote sensing data, particularly due to cloud cover and long revisit intervals, which may result in missing key seasonal observations and reduce classification accuracy.

All of the above contents have been added in Section 4.2 "Reliability of Global-PCG-10" of the manuscript (**see *Lines 459 ~ 484* for details**).

13. In lines 374-375, authors apply balanced sampling between PCG and non-PCG in the test set, which does not reflect their real-world scarcity and may lead to overestimated performance. Please discuss the impact of balanced sampling on evaluation or provide results under the true class distribution.

Response:

Thank you for this good question. Initially, we selected 20,500 test samples each for PCG and Non-PCG equally, with the primary goal of ensuring statistical stability for calculating the overall accuracy (OA) and evaluating the classification performance for the PCG category. However, through further literature review and methodological refinement, we have realized that this balanced sampling approach did not fully consider the effects of class imbalance on accuracy assessment.

To address this issue, we referred to the methodology proposed by Olofsson et al. (2014), and drew on best practices from land use and land cover classification studies such as Wang et al. (2023) and Tian et al. (2025) to re-sample PCG and Non-PCG and re-calculate the confusion matrix. In the updated method, we strictly followed the stratified random sampling strategy recommended by Olofsson et al. (2014), in which samples were selected in proportion to the actual mapped area of each class within the study region.

However, since PCG covers less than 1% of the global area, a strictly proportionate sampling approach would yield an insufficient number of PCG samples, making it difficult to effectively assess its classification accuracy. To address this issue, we adopted the approach used in the aforementioned studies and increased the proportion of PCG samples in the test dataset to approximately 10%. Now that the number of PCG is 6,000 while Non-PCG is 40,000. This adjustment could enhance the evaluation capability for this minority class (PCG) and ensures the scientific rigor and representativeness of the final accuracy estimates.

The above contents have been added in Section 4.2 "Reliability of Global-PCG-10" of the manuscript (**see *Lines 426 ~ 443* for details**).

References

Olofsson, P., Foody, G.M., Herold, M., Stehman, S.V., Woodcock, C.E. and Wulder, M.A., 2014. Good practices for estimating area and assessing accuracy of land change. *Remote sensing of Environment*, *148*, pp.42-57.

Tian, F., Wu, B., Zeng, H., Zhang, M., Zhu, W., Yan, N., Lu, Y. and Li, Y., 2025. GMIE: a global maximum irrigation extent and central pivot irrigation system dataset derived via irrigation performance during drought stress and deep learning methods. *Earth System Science Data*, *17*(3), pp.855-880.

Wang, M., Mao, D., Wang, Y., Xiao, X., Xiang, H., Feng, K., Luo, L., Jia, M., Song, K. and Wang, Z., 2023. Wetland mapping in East Asia by two-stage object-based Random Forest and hierarchical decision tree algorithms on Sentinel-1/2 images. *Remote Sensing of Environment*, *297*, p.113793.

14. In lines 381-382 the confusion matrix shows that the PCG class has a UA of 84.18% (FN = 3 243), yet no discussion of the false-negative causes is provided. Please analyze the reasons for these missed detections.

Response:

We sincerely thank you for the valuable comments. We apologize for the misunderstanding caused by the representation of the confusion matrix in the preprint version. The issue stemmed from an unclear explanation of the test sample construction. In the original dataset, 20,500 test samples were independently selected for each of the PCG and Non-PCG classes based on reference labels. However, in the preprint version, the confusion matrix failed to correctly reflect this setup due to misplacement of the values "2" and "3243", resulting in incorrect sample counts under each reference class.

The corrected confusion matrix is shown below.

Table 1. Confusion Matrix

Confusion Matrix	**Reference:** Non-PCG	**Reference:** PCG	*UA (%)*
Predicted: Non-PCG	20,498	3,243	86.34
Predicted: PCG	2	17,257	99.99
PA (%)	99.99	84.18	
F1-score (%)	-	91.41	
OA (%)			92.08

Note: PA, Producer's Accuracy; UA, User's Accuracy; OA, Overall Accuracy.*

Although there was a mistake in the early version of the confusion matrix, the accuracy metrics presented in Table 2 were correctly calculated from the actual test data. In the revised manuscript, this table has been removed according to Reviewer #1's suggestion.

In addition, to further assess the reliability of the Global-PCG-10 dataset from a quantitative perspective, we constructed an independent test sample set following the stratified random sampling strategy recommended by Olofsson et al. (2014), and consistent with practices from studies such as Wang et al. (2023) and Tian et al. (2025). Samples were drawn in proportion to the actual mapped area of PCG and non-PCG across the classification region.

The newly reconstructed confusion matrix is presented below, where "±" denotes the 95% confidence interval for each metric:

Table 1. Confusion Matrix

Confusion Matrix	**Reference:** Non-PCG Reference	**Reference:** PCG	*UA (%)*
Predicted: Non-PCG	39,991	893	97.82 ± 0.13
Predicted: PCG	9	5,107	99.82 ± 0.11
PA (%)	99.98 ± 0.01	85.12 ± 0.90	
F1-score (%)	-	91.88 ± 2.71	
OA (%)			98.04 ± 0.12

Note: PA, Producer's Accuracy; UA, User's Accuracy; OA, Overall Accuracy.*

In this result, FN = 893 and FP = 9. We have addressed the issue of false positives (FP = 9) in Response #12. The relatively high number of false negatives (FN = 893) can be attributed to the following factors. (1) Omission of small-scale PCG targets. Due to the 10-meter spatial resolution of Sentinel-2 imagery, which is significantly lower than that of high-resolution platforms like PlanetScope, small PCG often occupies only a few to a dozen pixels and are easily affected by mixed pixel issues. This makes it difficult for the model to extract reliable spectral features and leads to missed detections. (2) Limitations in spatiotemporal coverage of imagery. The Sentinel-2 data used in this study were organized by 1° grid tiles. Due to cloud contamination and observation scheduling constraints, it is sometimes challenging to obtain cloud-free imagery for both time periods (spring and summer), which reduces the model's ability to detect PCG in certain regions. (3) Post-classification filtering effects. To reduce false positives, we applied a strict post-processing procedure to the initial classification results when generating the Global-PCG-10 dataset. Specifically, a multi-stage Sieve Filter was used to remove small patches and isolated noise, which effectively suppressed misclassifications and significantly improved the precision (UA) for the PCG class.

Further analysis of the false negative (FN) cases, including field validation and image comparisons, will be presented in Section 4.5.2 Bad Case Analysis of the revised manuscript.

The above contents have been added in Section 4.2 "Reliability of Global-PCG-10" of the manuscript (**see *Lines 445 ~ 469* for details**).

References

Olofsson, P., Foody, G.M., Stehman, S.V. and Woodcock, C.E., 2013. Making better use of accuracy data in land change studies: Estimating accuracy and area and quantifying uncertainty using stratified estimation. *Remote sensing of environment*, *129*, pp.122-131.

Olofsson, P., Foody, G.M., Herold, M., Stehman, S.V., Woodcock, C.E. and Wulder, M.A., 2014. Good practices for estimating area and assessing accuracy of land change. *Remote sensing of Environment*, *148*, pp.42-57.

Tian, F., Wu, B., Zeng, H., Zhang, M., Zhu, W., Yan, N., Lu, Y. and Li, Y., 2025. GMIE: a global maximum irrigation extent and central pivot irrigation system dataset derived via irrigation performance during drought stress and deep learning methods. *Earth System Science Data*, *17*(3), pp.855-880.

Wang, M., Mao, D., Wang, Y., Xiao, X., Xiang, H., Feng, K., Luo, L., Jia, M., Song, K. and Wang, Z., 2023. Wetland mapping in East Asia by two-stage object-based Random Forest and hierarchical decision tree algorithms on Sentinel-1/2 images. *Remote Sensing of Environment*, *297*, p.113793.

15. In lines 381-382, non-PCG exhibits UA = 99.99% and PA = 86.30%, while PCG shows UA = 84.18% and PA = 99.99%, indicating a strong asymmetry in false-positive and false-negative distributions. Please discuss the origin of this bias.

Response:

In the reconstructed confusion matrix, the classification bias for the Non-PCG class has been substantially mitigated. However, the PCG class still exhibits an asymmetry in accuracy, i.e., its user's accuracy (UA) reaches as high as 99.99%, while the producer's accuracy (PA) remains relatively low at 84.18%. This discrepancy reflects an imbalance between false positives and false negatives in the model's predictions.

We believe this asymmetry is primarily caused by two factors. Firstly, under the 10-meter spatial resolution of Sentinel-2 imagery, many PCG areas consist of only a very limited number of pixels. In particular, small-scale PCG often occupies just a few, or a dozen mixed pixels. This makes it challenging for the model to extract robust spatial-spectral features, leading to frequent omissions and significantly lowering the PA for the PCG class. In Section 4.5.2, we further explore this issue through bad case analysis, where we provide visual examples of small PCGs being omitted. Secondly, to enhance the reliability of the final map product, we applied multiple post-processing steps to the Global-PCG-10 predictions, including a Sieve Filter to remove isolated or marginal misclassified regions. While this procedure effectively reduced false positives, thus improving the UA, it also suppressed small or fragmented PCG predictions, indirectly contributing to the lower PA.

In summary, this asymmetric accuracy reveals a systematic bias in the model's performance when

dealing with extremely rare, spatially small, and spectrally mixed targets such as PCG, rather than a result of random error. We elaborate on these issues in detail in the bad case analysis section **(*Lines 459 ~ 484*)** of the revised manuscript.

16. In section 4.3.2, authors compare visually against Tong et al. (2024) using 3 m data (Figure 12) but lack objective quantitative metrics. Please add quantitative evaluation and report specific comparative values.

Response:

Thank you for your constructive comment regarding the lack of objective quantitative metrics in Section 4.3.2. We have revised this section accordingly and taken the following measures to address your concern:

1. Accuracy assessment using 46,000 high-confidence test samples for Global-PCG-10

In this study, we conducted a quantitative evaluation of the Global-PCG-10 dataset using 46,000 high-confidence test samples, including 6,000 positive samples (PCG) and 40,000 negative samples (non-PCG). These samples were manually interpreted and verified to ensure high confidence. Specifically, the PCG samples were initially derived from the 2019 global 3-meter PCG map published by Tong et al. (2024), and were further validated using historical high-resolution imagery from Google Earth. However, upon comparing this validation sample set against both our Sentinel-2-based (10 m resolution) product and the 3-meter reference product (Tong et al., 2024), we observed that the 10 m product exhibited certain limitations in detecting small-scale PCG structures. This limitation is particularly pronounced for PCG areas that span only a few pixels and are often affected by mixed-pixel issues. This challenge likely contributes to the relatively lower user's accuracy (UA) for the PCG class in the confusion matrix results. To explain this phenomenon, we have added detailed quantitative analysis and bad case discussions in the revised Section 4.3.2.

2. Independent validation in high-density PCG regions and at the global scale

To provide a more objective and fair comparison, we followed the methodology proposed by Huang et al. (2022) and conducted a quantitative consistency analysis between the two datasets in terms of global PCG spatial distribution. Specifically, we selected four representative 1° × 1° grid regions with varying PCG densities. Each of these grids was further subdivided into multiple 0.01° × 0.01° sub-grid units. Within each sub-grid, we calculated the proportion of PCG pixels relative to the total number of pixels for both datasets (i.e., PCG area ratio, ranging from 0 to 1). Using these continuous ratio-based data, we applied linear regression analysis to calculate the coefficient of determination (R^2), thereby quantifying the spatial distribution consistency between the two datasets across different regions. Unlike methods that rely on discrete classification labels, this approach leverages continuous area proportions, making it more

suitable for evaluating agreement between remote sensing datasets with differing spatial resolutions. As shown in Figure 14a ~ d, the experimental results in four typical study area indicate that, in high-density PCG regions, our 10-meter resolution PCG dataset demonstrates a high degree of spatial consistency with the 3-meter reference dataset.

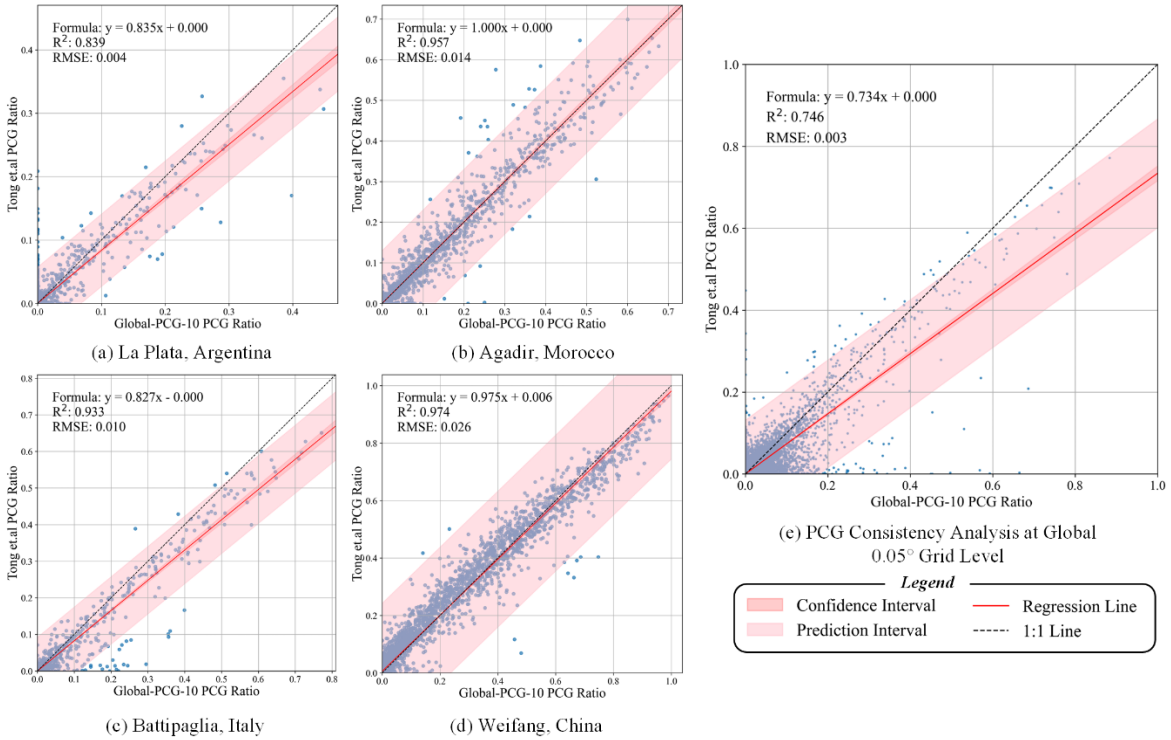

Figure 14. The consistency performance across the four representative regions and between the dataset by Tong et al. and Global-PCG-10 in representative regions.

To further evaluate spatial consistency at the global scale, we applied a standard regression-based consistency analysis across the entire globe, with reference to the analytical approach and spatial resolution (i.e., 0.05° grid) used by Huang et al, (2022). The coefficient of determination (R^2) was again employed as the primary evaluation metric. As shown in Figure 14e, the comparison based on a 0.05° grid reveals strong agreement in the global spatial distribution of PCG between the dataset published by Tong et al. (2024) and the Global-PCG-10 dataset. The regression analysis yields an R^2 of 0.746, a root mean square error (RMSE) of 0.003, and a regression equation of $y = 0.734x + 0.000$. These results indicate a moderate to strong spatial correlation between the two datasets, further validating the effectiveness of the Global-PCG-10 dataset in capturing the global distribution pattern of PCG.

As illustrated in Figure 14, the Global-PCG-10 dataset exhibits strong agreement with the reference data in typical regions (Figure 14a–d), whereas a moderate overestimation trend is observed at the global scale. This discrepancy may be attributed to the spatial resolution limitations of Sentinel-2 imagery. As a medium-resolution satellite (10m), Sentinel-2 is more susceptible to intra-class spectral variability and inter-class spectral confusion. In sparely distributed greenhouse areas, non-PCG features such as bare soil,

inter-greenhouse roads, or adjacent agricultural structures may exhibit spectral signatures similar to plastic-covered greenhouses, leading to misclassification and systematic overestimation of PCG coverage. Moreover, within the same spatial aggregation unit (e.g., a 0.05° grid cell), Sentinel-2 offers fewer pixels compared to PlanetScope (3m), making PCG area statistics more sensitive to per-pixel classification errors. Consequently, in typical regions with more homogeneous greenhouse patterns, clearer boundaries, the classification results are more stable and consistent. In contrast, at the global scale, the combined effects of landscape heterogeneity and resolution-induced error propagation contribute to reduced agreement.

The above contents have been added in Section 4.2 “Reliability of Global-PCG-10” of the manuscript (**see *Lines 551 ~ 572* for details**).

Reference:

Huang, X., Yang, J., Wang, W. and Liu, Z., 2022. Mapping 10-m global impervious surface area (GISA-10m) using multi-source geospatial data. *Earth System Science Data Discussions*, *2022*, pp.1-39.

17. Recommend including a discussion of potential applications and the limitations of the produced dataset.

Response:

We sincerely thank you for the valuable suggestion. We fully agree with your comment that further discussion is needed regarding the application potential and limitations of the proposed dataset to better illustrate its scientific value and boundaries of use.

Accordingly, we have revised and expanded Section 4.4 in Chapter 4 of the manuscript to include the following subsection.

4.4 Application potential and limitations of the dataset

As described above, Global-PCG-10 is a global-scale dataset of PCG derived from open-access Sentinel-2 imagery. By leveraging freely available satellite data, the dataset significantly reduces production costs while providing a standardized and well-structured data format that can be easily integrated with other open-source remote sensing products.

As the first global PCG dataset with 10-meter spatial resolution, Global-PCG-10 has strong application potential in various domains. (1) In agricultural monitoring and statistics, the dataset reveals the spatial distribution pattern of global protected agriculture, offering valuable support for agricultural structure optimization, farmland use monitoring and irrigation estimation. (2) In agro-environmental assessments, it provides high-resolution spatial information on protected agriculture, supporting efforts

by governments and international organizations to conduct agricultural censuses, develop regional agricultural strategies and implement climate-adaptive agricultural policies. (3) In open-source land use/land cover (LULC) applications, PCG are often underrepresented in current global LULC products. This dataset helps fill that gap by explicitly including PCG as a key cropland subtype.

Despite its usefulness, Global-PCG-10 still has several limitations that need to be addressed in future work. Firstly, due to the 10-meter resolution of Sentinel-2 imagery, it remains difficult to detect small-scale or scattered PCG units, especially in regions dominated by smallholder agriculture. This may lead to omission errors. In the future, we plan to integrate higher-resolution remote sensing data to develop regional PCG datasets with finer spatial detail. Secondly, the classification task in this study focused primarily on the overall category of PCG, without further distinguishing among its subtypes. In future research, we plan to explore fine-grained classification methods for agricultural greenhouses (AG), including the differentiation of daylight greenhouses, conventional plastic greenhouses and small arch sheds, in order to further enhance the accuracy and practical applicability of PCG dataset. Thirdly, as the dataset only contains PCG in 2020, it does not capture dynamic PCG changes such as recent expansion or degradation regions. We plan to extend this work to develop a global time-series dataset of greenhouses, enabling long-term monitoring and trend analysis. Besides, the current pipeline for PCG mapping, which combines deep learning and active learning, still relies on a semi-automated weak-label updating strategy and does not yet support full end-to-end automation. In the future, we aim to explore end-to-end weak-label learning frameworks to build a more efficient and automated data processing system.

The above contents have been added in Section 4.2 “Application potential and limitations of the dataset” of the manuscript (**see *Lines 573 ~ 597* for details**).

Thank you again for your comments. They are valuable and very helpful for revising and improving our paper, as well as the important guiding significance to our studies.

Yours sincerely,

Bowen Niu, Quanlong Feng

on behalf of all the co-authors