Dear Editor and Reviewers,

Manuscript ID essd-2026-24 entitled "GlobalRice20: A 20 m resolution global paddy rice dataset for 2015 and 2024 derived from multi-source remote sensing."

We would like to express our sincere gratitude to the Editor and all four reviewers for their constructive feedback and thorough review of our manuscript. We have carefully considered all of the comments and have made corresponding revisions throughout the manuscript. In addition, following the reviewers' suggestions, we have polished the English expressions throughout the manuscript and improved the clarity of several figures. Below, we provide detailed point-by-point responses to all comments raised by the four reviewers, together with corresponding revisions and clarifications where necessary. We sincerely hope that these revisions have adequately addressed the concerns and uncertainties raised by the reviewers.

In the manuscript and this response letter, revisions made in response to Reviewer 1 are highlighted in red, those in response to Reviewer 2 are highlighted in green, those in response to Community Comment 1 (Dr. Ran Huang) are highlighted in orange, and those in response to Community Comment 2 (Dr. Dailiang Peng) are highlighted in blue. In addition, to improve the readability and overall quality of the paper, additional modifications are highlighted in cyan.

Once again, we sincerely thank you for your time and effort in reviewing our work, and we look forward to your feedback.

Sincerely,

Zhang Hong

zhanghong@radi.ac.cn

# Response to Reviewer 1

**Comments to the Author**

**This paper presents GlobalRice20, claimed to be the first global 20 m resolution paddy rice dataset covering 98 countries for 2015 and 2024. The core methodological contribution is T2VRCM, a framework that converts irregular, multi-source remote sensing time series (Sentinel-1 VV/VH, and optical EVI/LSWI from Sentinel-2 or Landsat-8) into 2D line-graph images, which are then classified by a custom CNN with gated convolution blocks and a full-stage feature interaction module. The model achieves 92.33% overall accuracy on a 164,000-sample test set and shows $R^2 = 0.91$ agreement with national statistics for 2024. Decadal analysis reveals a 6.6% global rice area increase, with Africa showing the strongest growth at 15.7%.**

**The topic is timely and the ambition to produce a global 20 m rice map is commendable. Overall, this dataset would be a valuable resource to the community. However, before this paper can be accepted, I have the following concerns for authors to address.**

**RESPONSE:** We sincerely thank you for your careful evaluation of our manuscript and for recognizing the potential scientific and practical value of GlobalRice20. Your detailed and constructive comments have been extremely valuable in helping us further improve the scientific rigor, completeness, and clarity of the manuscript. We have carefully examined each of your concerns and provided detailed responses to all comments below.

**1. The "stable sample" strategy introduces selection bias. Validation samples are restricted to locations that maintained consistent land cover across both 2015 and 2024. This systematically excludes the most challenging pixels, areas where land use changed, where rice fields were abandoned or newly established, or where**

**confusion with other crops is most likely. The reported 92.33% OA therefore overestimates real-world performance. Accuracy on dynamic areas, which are precisely where the dataset would be most useful for SDG 2 monitoring, remains unknown. At minimum, additional validation on independently sampled points (including changed areas) is needed.**

**RESPONSE:** Thank you for your valuable feedback. We entirely agree with your assessment; while the "stable sample" strategy provided reliable, low-noise labels for model training, it excluded pixels where land cover changed during the validation phase. Therefore, the reported 92.33% overall accuracy (OA) based solely on stable samples likely overestimates the true, overall performance in real-world scenarios that include dynamic changes. In response to your comments, we have supplemented our work with the following:

First, we constructed a supplementary dynamic change test set based on the existing sample set. To optimize sampling efficiency, we utilized the temporal differences between the 2015 and 2024 GlobalRice20 mapping results to delineate candidate regions globally where land cover changes likely occurred. Using countries as the spatial stratification unit, we performed spatially stratified random sampling to extract 16,400 sample points for each target year. We ensured a balanced representation between two primary change trajectories: newly established rice fields (non-rice changing to rice) and abandoned rice fields (rice changing to non-rice). All candidate points were collaboratively labeled by a dedicated annotation team comprising 8 graduate students, 4 professors with agricultural remote sensing expertise, and 1 expert from the Academy of Agricultural Sciences. Guided by standardized visual interpretation rules established by the expert, the 8 graduate students independently determined the true land cover classes for 2015 and 2024 using the corresponding optical imagery. Subsequently, the 4 professors comprehensively audited the annotation results, resolving any ambiguities through field investigations and team consensus meetings. This process was entirely independent of the model's own predictions. The labeling criteria remained strictly consistent with the original 164,000 stable samples.

To illustrate the dynamic changes recorded in the newly added test set, we have plotted Figure R1, which demonstrates the transitions in land cover types across the two observation years. For instance, Position 1 highlights an area transitioning from active rice cultivation in 2015 to non-rice in 2024. Conversely, Positions 2 and 3 depict the establishment of new rice paddies.

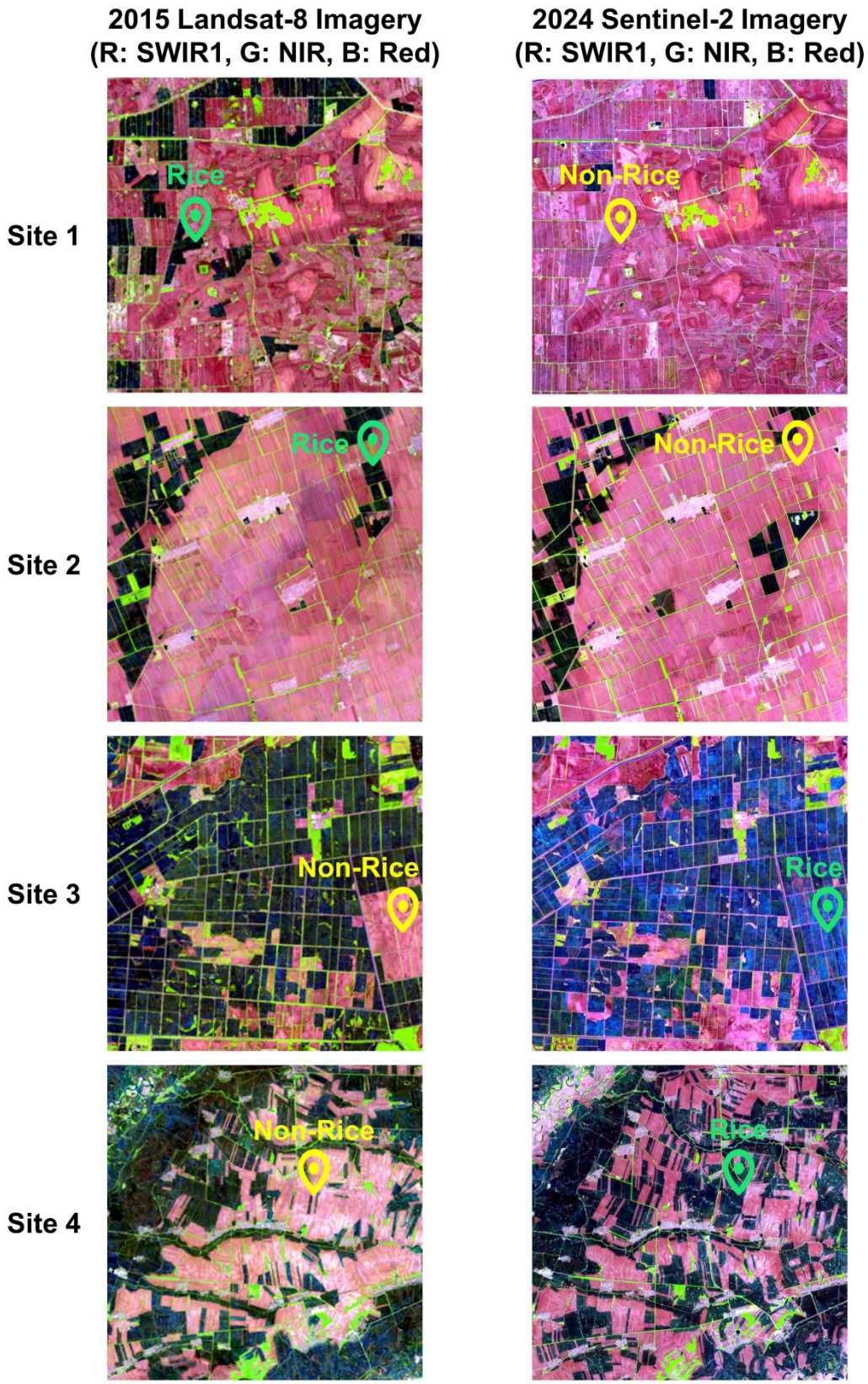


**Figure R1. Examples of dynamic land cover transitions within the supplementary validation set. The left column displays 2015 Landsat-8 imagery, and the right column displays 2024 Sentinel-2 imagery.**

The newly generated dynamic test set serves as a completely independent subset; it did not participate in any model training and was exclusively used to evaluate the trained T2VRCM models. We utilized this dynamic test set to validate the accuracy of the rice distribution products for both years, and the results are summarized in Table R1.

**Table R1. Accuracy comparison across stable, dynamic, and combined test subsets for 2015 and 2024.**

| Year | Test Subset | Sample Size | OA (%) | PA (%) | UA (%) | F1 (%) |
|---|---|---|---|---|---|---|
| 2015 | Stable Test Set | 32,800 | 90.23 | 90.37 | 90.11 | 90.24 |
| 2015 | Dynamic Test Set | 16,400 | 88.37 | 86.78 | 89.62 | 88.18 |
| 2015 | Combined | 49,200 | 89.61 | 89.17 | 89.96 | 89.56 |
| 2024 | Stable Test Set | 32,800 | 92.33 | 92.12 | 92.52 | 92.32 |
| 2024 | Dynamic Test Set | 16,400 | 90.35 | 88.68 | 91.75 | 90.19 |
| 2024 | Combined | 49,200 | 91.67 | 90.98 | 92.27 | 91.62 |

The results show that the OA of the dynamic test set for 2024 is 90.35%, which is lower than that of the stable regions. Errors were primarily concentrated around newly established paddies and in transitional zones where rice intermingles with wetlands or other crops. The combined OA of the stable and dynamic test sets for 2024 is 91.67%. For 2015, the corresponding OAs for the stable, dynamic, and combined test sets are 90.23%, 88.37%, and 89.61%, respectively. The results indicate that the combined overall accuracy remains at a high level, demonstrating that the T2VRCM model retains strong generalization capabilities in real-world scenarios that include dynamic changes.

In summary, we believe that this supplementary experiment involving the dynamic change test set not only upholds the core conclusions of our original accuracy assessment but also provides a direct and specific response to your concern regarding the unknown real-world performance in dynamic areas. Thank you once again for

prompting us to conduct a more comprehensive and rigorous accuracy validation of our dataset.

**2. The 2015 optical data come from Landsat-8 at 30 m, and Sentinel-1 GRD has a pixel spacing of 10 m but an effective spatial resolution closer to 20 m in range. Resampling 30 m Landsat pixels to 20 m does not create information at that finer scale. The paper should clearly state that the 2015 product has a coarser effective resolution than the 2024 product and discuss the implications for change detection between the two years. Calling both "20 m resolution" without this caveat is misleading.**

**RESPONSE:** Thank you very much for raising this important point. We agree with your observation that the optical data used in 2015 were from Landsat-8 OLI, which has a native spatial resolution of 30 m. Meanwhile, although Sentinel-1 GRD data have a 10 m pixel spacing, their range spatial resolution is approximately 20 m. We also agree that resampling the 30 m Landsat pixels to a 20 m grid does not, by itself, introduce additional fine-scale optical spatial information.

We would like to clarify that the term "20 m resolution" used in the manuscript refers to the standardized output grid and spatial resolution of the final GlobalRice20 product, rather than implying that all input datasets had exactly the same native spatial resolution in both years. We have provided a more detailed description of the data processing procedures, including resampling, in Section 2.2.

The 2015 product was not generated solely by resampling Landsat-8 optical data. Instead, Landsat-8 optical indices and Sentinel-1 VV/VH SAR time series were jointly used for classification. The Sentinel-1 SAR data provide a 10 m pixel spacing and an effective spatial resolution of approximately 20 m, thereby contributing higher-resolution spatial information as well as important phenological information for rice identification in 2015, particularly in rice-growing regions where cloud and rainfall conditions resulted in insufficient optical observations. Therefore, we believe that

publishing the final product on a standardized 20 m grid is reasonable. Nevertheless, we fully agree that this description should be further clarified.

Since the use of Landsat-8 data in 2015 is one of the factors contributing to the slightly lower accuracy of the 2015 product compared with the 2024 product, we have added a corresponding explanation in Section 4.3.3. In addition, the 2015–2024 change analysis was not conducted as a strict pixel-by-pixel change detection at the 20 m pixel scale. Instead, large-scale changes in global rice cultivation patterns were characterized through area aggregation at a 1° grid scale. Conducting the change analysis after aggregation to 1° grid cells helps reduce the influence of differences between sensors on the observed trends in rice area changes. We have added a clarification of this point in the revised manuscript. Once again, we sincerely appreciate your valuable comments.

The specific revisions are as follows.

**Pages 4-5**

We selected Interferometric Wide (IW) swath mode Ground Range Detected (GRD) products with a nominal pixel spacing of 10 m.

Both VV and VH polarizations were utilized, and ascending and descending acquisitions were aggregated together into 12-day mean composites after local incidence-angle normalization to reduce viewing-geometry-related backscatter differences (Kaplan et al., 2021; Arias et al., 2022).

Finally, all SAR and optical data were resampled to a common 20-m spatial grid to ensure consistent spatial alignment.

**Line 205, Page 25**

The slightly lower accuracy in 2015 can mainly be attributed to the longer revisit cycle and 30 m spatial resolution of Landsat-8 optical imagery, together with the relatively limited SAR coverage in that year. These factors weakened the representation of time-series image features in some regions and may have reduced the delineation of fine field boundaries compared with the 2024 Sentinel-2-based product, thereby affecting the

stability and accuracy of area estimation, as reflected in the overestimation of rice area in India in 2015.

**Page 27**

Global rice area was aggregated into a 1° grid to calculate the area change between 2015 and 2024 thereby reducing the influence of local boundary differences associated with sensor resolution and data availability on the interannual comparison (Figure 10).

**3. The comparison with native time-series models is unfair. For RF, LSTM, and Transformer baselines, missing temporal phases were filled by linear interpolation and completely absent modalities were zero-filled. Zero-filling is a known poor strategy since it conflates "missing data" with "zero signal", which is especially harmful for SAR backscatter where zero is not a physically meaningful value. More appropriate strategies exist (e.g., learnable masking tokens, attention masks, multi-task imputation). The paper should either apply state-of-the-art missing-data handling for the time-series baselines or acknowledge this limitation explicitly. Without this, the claim that the time-series-to-vision strategy is inherently superior is not well supported.**

**RESPONSE:** Thank you very much for this valuable comment. We apologize that the original manuscript did not clearly distinguish between different types of missing data and their corresponding handling strategies. In particular, partial temporal missingness and complete modality absence were described together, which may have obscured the differences between these two cases and led to misunderstanding of the experimental settings. Following your suggestion, we carefully re-examined the relevant experimental settings and have further clarified the corresponding procedures and supplemented the analysis.

First, we would like to clarify that two fundamentally different types of missing data are encountered in this study: partial temporal missingness and complete modality absence. These two cases are handled differently.

The first case is partial temporal missingness, for which zero filling was not used in this study. For the optical time series, missing observations mainly result from invalid acquisitions caused by clouds and cloud shadows. For the Sentinel-1 SAR time series, some temporal observations may also be unavailable in certain regions. In these cases, valid observations remain available before and after the missing temporal positions, allowing the missing values to be interpolated using neighboring valid observations. For RF, LSTM, and Transformer models, which require inputs with fixed temporal dimensions, linear interpolation was therefore used to construct regularized time series.

We agree with the reviewer that more advanced approaches, including learnable imputation, joint reconstruction, explicit missing-data masks, and attention masks, have increasingly been used for modeling irregular time series and may further improve the ability of native time-series models to handle missing observations. However, when extended to large-scale global mapping, such approaches often require the construction of training data that reflect different missing-data patterns, together with additional model training and parameter optimization, thereby increasing computational demand and implementation complexity. Because this study performs a unified large-scale mapping task across 98 countries, we adopted linear interpolation for the baseline experiments because it provides a stable and computationally efficient solution without requiring an additional imputation model. This choice also helps maintain the operational feasibility of the global processing workflow and a consistent comparison framework across baseline models. Following the reviewer's suggestion, we have clarified this preprocessing procedure in the revised manuscript and explicitly acknowledged in the limitations that more advanced masking- and learning-based missing-data handling strategies may further improve the performance of native time-series models.

The second case is complete modality absence, which differs fundamentally from partial temporal missingness in both its cause and the corresponding processing strategy. This situation mainly occurs in some regions where, because of the actual spatial coverage of Sentinel-1 acquisitions, no valid Sentinel-1 observations are available

during the target year. In such cases, the entire annual time series of both VV and VH is absent. Because no valid observations exist within the missing modality, the corresponding SAR time series cannot be reconstructed through temporal interpolation as in the case of partial temporal missingness.

For RF, LSTM, and Transformer models, which require fixed-dimensional numerical inputs, the original experiments used a constant value of 0 to construct placeholder sequences for the completely missing SAR modality. We would like to emphasize that 0 was not intended as a physical estimate of missing SAR backscatter, but only as a numerical placeholder required to maintain the fixed input dimensionality of these models. To examine whether the choice of a specific placeholder value affected model performance and to further assess the robustness of the results, we conducted an additional sensitivity analysis for cases of complete SAR modality absence. In addition to the baseline value of 0, we tested fixed values of −999 and 999, as well as the training-set minimum, maximum, mean, and median as placeholder values for the completely missing SAR modality. RF, LSTM, and Transformer were then re-evaluated using the same data split, training configuration, and model settings. The results are shown below.

**Table R2. Sensitivity analysis of different numerical placeholders for complete SAR modality absence.**

| SAR modality placeholder | RF F1 (%) | LSTM F1 (%) | Transformer F1 (%) |
|---|---|---|---|
| Fixed 0 (baseline) | 82.87 | 84.79 | 85.45 |
| Fixed −999 | 82.86 | 84.78 | 85.44 |
| Fixed 999 | 82.86 | 84.78 | 85.44 |
| Training-set minimum | 82.87 | 84.79 | 85.45 |
| Training-set maximum | 82.87 | 84.79 | 85.45 |
| Training-set mean | 82.85 | 84.78 | 85.44 |

| SAR modality placeholder | RF F1 (%) | LSTM F1 (%) | Transformer F1 (%) |
|---|---|---|---|
| Training-set median | 82.85 | 84.78 | 85.44 |

The results show that the classification performance of all three native time-series baselines, namely RF, LSTM, and Transformer, remained essentially unchanged across placeholder values with substantially different numerical ranges. Although slight performance fluctuations were observed, the maximum variation in F1 was only 0.02 percentage points for RF and 0.01 percentage points for both LSTM and Transformer. Whether fixed extreme values (−999 and 999), the training-set minimum and maximum, or the training-set mean and median were used, the choice of placeholder had no substantive effect on classification accuracy.

These two types of missing data also constitute an important motivation for the proposed Time-Series-to-Vision (T2V) representation. For partial temporal missingness, T2V does not require numerical imputation of missing observations onto a regular temporal grid before representation. Instead, the available observations are organized and connected according to their actual Day of Year (DOY) positions, thereby preserving the temporal information contained in the available observations as much as possible. When Sentinel-1 SAR data are completely unavailable throughout the year, the corresponding SAR subplot is left blank, explicitly representing the unavailability of that modality without artificially constructing a SAR time series that does not exist. In such cases, T2VRCM primarily relies on the optical temporal features that remain available for classification.

Therefore, from the perspective of the actual remote-sensing information available to the models, neither T2V nor RF, LSTM, and Transformer receives additional SAR observations for samples with complete SAR modality absence. T2V explicitly represents the missing modality using a blank subplot, whereas fixed-dimensional time-series models require a numerical placeholder to maintain the corresponding input positions. The additional sensitivity analysis further demonstrates that the classification

performance of the native time-series models does not depend on the specific use of 0 as the placeholder value.

The overall results reported in Table 1 provide further supporting evidence. Among the general-purpose vision classification models using the same T2V representation, even ResNet-18, which achieved the lowest F1 among these vision models, reached an F1 of 87.73%. This is 4.86 percentage points higher than RF (82.87%) and 2.28 percentage points higher than Transformer, the best-performing native time-series baseline (85.45%). Under the same T2V representation, ViT-Tiny, Swin-Tiny, ConvNeXt-Atto, and MambaOut-Femto achieved F1 values of 88.25%, 88.59%, 88.70%, and 89.36%, respectively, all exceeding those of the compared RF, LSTM, and Transformer models. Furthermore, T2VRCM, which was specifically designed for the T2V representation, achieved an F1 of 92.32%, exceeding RF, LSTM, and Transformer by 9.45, 7.53, and 6.87 percentage points, respectively.

Taken together, these results indicate that the performance differences observed between the native time-series models and the vision-based models in the current experiments cannot be simply attributed to the use of 0 as the placeholder for complete SAR modality absence. More importantly, when the same T2V representation was used, multiple vision architectures with substantially different network designs consistently achieved higher classification accuracy than the compared native time-series models, while T2VRCM further improved performance beyond these general-purpose vision models. These results demonstrate that, under the global multi-source remote-sensing scenarios considered in this study, which involve irregular observations and modality absence, transforming time series into a unified two-dimensional visual representation provides a suitable and effective representation strategy.

At the same time, we fully agree with the reviewer that more advanced missing-data modeling strategies, including learnable masking, attention masks, missingness indicators, and learning-based imputation or reconstruction, have the potential to further improve the ability of time-series models to handle missing observations and enhance classification performance. We greatly appreciate this insightful suggestion

and have added a corresponding discussion to the revised manuscript. Future work will further investigate the applicability of these approaches to global-scale remote-sensing time-series classification.

We sincerely thank the reviewer again for this constructive comment.

The specific revisions are as follows.

**Page 16**

For cases of complete SAR modality absence, zero was used solely as a numerical placeholder to satisfy the fixed-dimensional input requirement rather than as a physical estimate of the missing SAR backscatter.

**Page 32**

Third, for the native time-series baseline models, linear interpolation was adopted to regularize incomplete temporal observations. This strategy provides a stable and computationally efficient solution for large-scale experiments, but it may not fully capture complex patterns of missing observations. More advanced temporal interpolation and missing-data handling approaches may further improve the ability of time-series models to process irregular and incomplete observations. Evaluating the effectiveness and scalability of these approaches in global crop-mapping applications therefore represents an important direction for future research.

**4. Train/test splitting protocol is insufficiently described. The paper states 164,000 samples across 98 countries but does not specify: (1) the ratio of training to testing, (2) whether the split is spatially stratified to prevent spatial autocorrelation leakage, and (3) whether the same samples are used for both 2015 and 2024 evaluation (which they appear to be, given the "stable sample" constraint). If nearby pixels appear in both training and test sets, accuracy will be inflated. Readers who wish to reuse or benchmark against this dataset need these details.**

**RESPONSE:** Thank you for your valuable feedback. We fully agree that clearly

articulating the training/test set partitioning strategy, particularly the split ratio, spatial stratification approach, and cross-year sample reuse, is crucial for the credibility and reproducibility of the results, as well as for future benchmarking of the dataset. In response to your three specific questions, our clarifications are as follows, and the relevant details have been incorporated into Section 2.3.1 of the revised manuscript:

(1) Regarding the training-to-test set ratio

In this study, a total of 164,000 stable samples were constructed (131,200 for training and 32,800 for testing), corresponding to a training-to-test ratio of 4:1. This ratio and allocation strategy perfectly align with common practices in similar large-scale mapping efforts within the remote sensing domain (Han et al., 2021; Song et al., 2025; Jiang et al., 2024).

(2) Regarding the spatial stratification strategy to mitigate spatial autocorrelation leakage

We employed a spatially stratified random sampling strategy, utilizing each country as a spatial stratification unit. Within each country, samples were independently and randomly partitioned into the training and test sets, strictly maintaining the 4:1 ratio. Simultaneously, we ensured that the 1:1 balance between rice and non-rice samples within each country was preserved in both the training and test sets. This country-based stratification guarantees that the model is adequately and evenly trained and validated across different climate zones and cropping systems. Furthermore, all sample points were independently selected through manual visual interpretation of discrete locations distributed across rice-cultivating and non-cultivating regions in each country, rather than being sampled from continuous spatial blocks. The substantial spatial intervals generally maintained between sample points mitigate the risk of adjacent or duplicate pixels appearing in both the training and test sets, thereby alleviating the issue of accuracy overestimation driven by spatial autocorrelation. This methodological approach is consistent with sample partitioning practices in related studies (Xu et al., 2023; Zhao et al., 2024).

(3) Regarding the use of a shared sample set for 2015 and 2024

As you correctly deduced, based on the "stable sample" strategy described in Section 2.3.1, both 2015 and 2024 share the exact same set of geographically fixed sample points (i.e., locations that maintained a consistent land cover class across both years). It is important to note, however, that we independently trained two separate T2VRCM models for 2015 and 2024. Each model utilized only the remote sensing image time-series features corresponding to its specific year for training and testing. Both training and accuracy assessment were strictly confined to the contemporary training and test subsets of that respective year; there is no cross-year or cross-stage time-series reuse or data leakage. Additionally, the assignment of a sample point at a specific location to either the training or test set remained constant across both years. This ensured a fair comparison between the two T2VRCM models based on an identical sample foundation.

In summary, we have added detailed explanations regarding the three aforementioned aspects to Section 2.3.1 of the revised manuscript. This will facilitate readers' understanding and reproduction of our dataset's partitioning methodology while providing essential information for future benchmarking. Thank you once again for this important and constructive comment, which has significantly enhanced the transparency of the methodology section in our paper.

The specific revisions are as follows.

**Page 6-7**

**2.3.1 Reference Sample Set**

To train and validate the T2VRCM model, a comprehensive reference dataset was constructed containing 164,000 samples across 98 countries. All samples were derived through rigorous visual interpretation of discrete locations distributed across rice-cultivating regions in each country, rather than being sampled from continuous spatial blocks. Consequently, this design mitigates the likelihood of adjacent or duplicate samples appearing in both the training and test sets, thereby reducing the risk of accuracy overestimation driven by spatial autocorrelation.

The sample annotation primarily relied on high-resolution optical imagery corresponding to the target years (2015 and 2024), including Google Earth, Landsat-8, and Sentinel-2 imagery. Additionally, existing regional rice maps were incorporated to locate potential cultivation extents. The annotation process was collaboratively executed by a dedicated team comprising 8 graduate students, 4 professors with extensive experience in agricultural remote sensing, and 1 expert from the Academy of Agricultural Sciences. This process involved multiple stages: initially, the expert provided unified guidance and established standardized visual interpretation rules for the team; subsequently, the 8 graduate students conducted preliminary labeling of candidate samples based on these criteria; finally, the 4 professors comprehensively audited the preliminary results. Any discrepancies encountered during the review were resolved through multiple team consensus meetings, thereby ensuring the high accuracy and reliability of the final labels.

The sample points were partitioned into training and test sets employing a spatially stratified random sampling strategy, utilizing each country as a stratification unit. Within each country, samples were allocated to the training and test sets at a fixed ratio of 4:1 (yielding approximately 131,200 training samples and 32,800 test samples globally). A balanced 1:1 ratio between rice and non-rice samples was maintained in both subsets to prevent class imbalance.

Crucially, to ensure temporal consistency and reduce label noise, we adopted a "stable sample" strategy: selected samples were restricted to locations that maintained a consistent land cover type (either persistent rice or persistent non-rice) in both 2015 and 2024. Consequently, identical sample locations and training/test splits were applied for both 2015 and 2024. However, we independently trained and evaluated two separate T2VRCM models for the two target years, utilizing the time-series features corresponding to each respective year. The training and evaluation of each model were strictly confined to the contemporary training and test sets of that specific year.

**References**

Han, J., Zhang, Z., Luo, Y., Cao, J., Zhang, L., Cheng, F., Zhuang, H., Zhang, J., and Tao, F.: NESEA-Rice10: high-resolution annual paddy rice maps for Northeast and Southeast Asia from 2017 to 2019, Earth system science data, 13, 5969-5986, 2021.

Jiang, J., Zhang, H., Ge, J., Zuo, L., Xu, L., Song, M., Ding, Y., Xie, Y., and Huang, W.: 20 m africa rice distribution map of 2023, Earth System Science Data Discussions, 2024, 1-34, 2024.

Song, M., Xu, L., Ge, J., Zhang, H., Zuo, L., Jiang, J., Ding, Y., Xie, Y., and Wu, F.: EARice10: a 10 m resolution annual rice distribution map of East Asia for 2023, Earth system science data, 17, 661-683, 2025.

Xu, S., Zhu, X., Chen, J., Zhu, X., Duan, M., Qiu, B., Wan, L., Tan, X., Xu, Y. N., and Cao, R.: A robust index to extract paddy fields in cloudy regions from SAR time series, Remote Sensing of Environment, 285, 113374, 2023.

Zhao, Z., Dong, J., Zhang, G., Yang, J., Liu, R., Wu, B., and Xiao, X.: Improved phenology-based rice mapping algorithm by integrating optical and radar data, Remote Sensing of Environment, 315, 114460, 2024.

**5. Country-level $R^2$ conflates spatial accuracy with area estimation. A map with spatially compensating errors (overestimation in one region, underestimation in another) can still achieve a high $R^2$ against national statistics. This metric does not validate spatial accuracy at the pixel or field level. The paper should acknowledge this limitation. Subnational (e.g., province-level) comparisons, at least for countries where such statistics are available (China, India, USA), would strengthen the validation considerably.**

**RESPONSE:** We thank the reviewer for this valuable suggestion. We agree that the national-scale $R^2$ mainly reflects the overall agreement between GlobalRice20-derived rice areas and official statistical areas, and does not directly represent pixel-level or field-level spatial accuracy. Because local overestimation and underestimation within a

country may offset each other, relying solely on national-scale statistical comparisons could indeed mask some spatial errors. Therefore, following the reviewer's suggestion, we added subnational-scale statistical comparisons to further strengthen the validation of the area estimates. Specifically, we selected representative countries with relatively complete rice area statistics at the first-level administrative-unit scale, including China, India, and the United States. Rice areas extracted from GlobalRice20 were aggregated to provincial or state-level administrative units and compared with the corresponding official statistics. The provincial rice planting area data for China were obtained from the China Rural Statistical Yearbook published by the National Bureau of Statistics of China. State-level rice area data for the United States were obtained from the Rice Yearbook provided by the USDA Economic Research Service. State-level rice area data for India were obtained from the Handbook of Statistics on Indian States published by the Reserve Bank of India. We have added descriptions of these subnational statistical data sources in Section 2.3.2 of the revised manuscript and included the corresponding provincial/state-level agreement validation results in Section 4.3.3. This additional validation helps reduce the influence of offsetting spatial errors in national-scale statistics and therefore provides more rigorous support for the reliability of GlobalRice20 area estimates at regional scales. Once again, we sincerely appreciate your valuable suggestion.

The specific revisions are as follows:

**Page 8**

Furthermore, this study selected China, the United States, and India, three major rice-producing countries, to conduct validation analysis at the first-level administrative division scale. The rice area data used in this study were obtained from national statistical sources. The provincial rice cultivation area data for China were derived from the 2015 and 2024 editions of the China Rural Statistical Yearbook, compiled and published by the National Bureau of Statistics of China (NBS) (http://www.tjcn.org). The state-level rice cultivation area data for the United States were obtained from the Rice Yearbook provided by the United States Department of Agriculture, Economic

Research Service (USDA ERS) (https://www.ers.usda.gov/data-products/rice-yearbook). The state-level rice cultivation area data for India were obtained from the Handbook of Statistics on Indian States published by the Reserve Bank of India (RBI) (https://www.rbi.org.in/Scripts/AnnualPublications.aspx?head=Handbook+of+Statistics+on+Indian+States).

**Page 25-26**

To further evaluate the consistency between GlobalRice20-derived rice areas and official statistics at a finer spatial scale, we conducted a subnational validation using first-level administrative units in China, the United States, and India (Fig. 9). The results showed strong agreements between the estimated and statistical rice areas across all three countries. Specifically, the $R^2$ values for China reached 0.88 in 2015 and increased to 0.93 in 2024. The United States exhibited the highest consistency, with $R^2$ values of 0.98 and 0.99 in 2015 and 2024, respectively. For India, the $R^2$ values were 0.90 in 2015 and 0.93 in 2024. Overall, the subnational validation results showed higher consistency in 2024 than in 2015 across all three countries. This improvement is likely attributable to the enhanced spatial and temporal observation capabilities of Sentinel-2 imagery compared with Landsat-8, which provided finer spatial resolution and more frequent observations for capturing rice cultivation dynamics. Among the three countries, the United States achieved the highest agreement between mapped and statistical rice areas. This may be related to the relatively regular spatial distribution of rice fields, larger field sizes, and more homogeneous cropping systems, which reduce the complexity of rice boundary delineation and classification. In comparison, China and India contain more diverse rice cultivation landscapes, including fragmented smallholder fields, complex cropping patterns, and greater regional variability, which introduce additional challenges for accurate area estimation. Nevertheless, the consistently high correlations across all three countries demonstrate the reliability of GlobalRice20 for regional-scale rice area estimation.

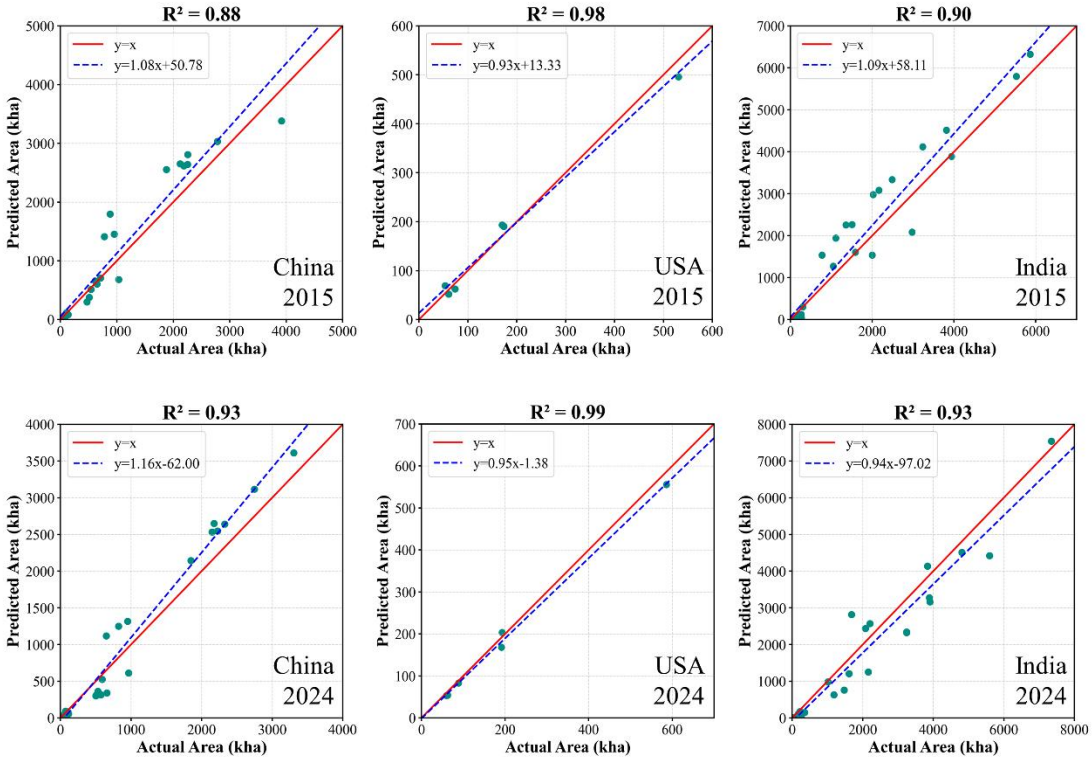


**Figure 9. Subnational-scale validation of GlobalRice20 against official rice statistics.**

**6. The paper jumps from Results (Section 4) directly to Conclusion (Section 5). A dedicated Discussion section is expected in ESSD to address: limitations of the method and dataset, potential error sources (e.g., confusion with other flooded crops, aquaculture, seasonal wetlands), sensitivity to SAR coverage gaps, and the impact of using different optical sensors across the two years. The absence of any honest discussion of failure cases or systematic errors is a significant gap.**

**RESPONSE:** Thank you very much for this valuable comment. We agree that a systematic discussion of the limitations of the proposed method and dataset, the effects of different data conditions on classification performance, and the potential sources of error is important for providing a more comprehensive and objective assessment of GlobalRice20. Following your suggestion, we added a dedicated Discussion section to the revised manuscript and further analyzed interannual data configurations, Sentinel-1 data coverage, as well as the limitations and potential error sources of the dataset.

First, to examine the effects of using different optical sensors and SAR data conditions in 2015 and 2024, we conducted modality ablation experiments separately for the two years. Three input configurations, namely Optical-only, SAR-only, and Optical+SAR, were compared to quantify the effects of different optical and SAR data configurations on rice classification performance. Second, to further evaluate the influence of Sentinel-1 data coverage gaps, we conducted a coverage sensitivity analysis using the independent test samples from 2024. The samples were divided into No coverage, Partial coverage, and Complete coverage according to the availability of Sentinel-1 observations during the year, allowing us to systematically compare model performance under different SAR coverage conditions. These analyses have been incorporated into Discussion Sections 5.1 and 5.2, respectively.

In addition, we further analyzed the potential causes of local misclassification in GlobalRice20 and provided representative examples involving aquaculture areas, wetlands, and flooded vegetation, as shown in Figure R2. These land-cover types may exhibit inundation characteristics, spectral responses, SAR backscatter patterns, and vegetation dynamics similar to those of paddy rice during certain periods, which can lead to local confusion. Regional differences in rice cropping calendars and field management practices, mixed pixels along field boundaries, persistent cloud contamination, incomplete Sentinel-1 coverage, and differences in the spatial and temporal resolutions of the input imagery between the two years may further increase classification uncertainty. Although the fusion of optical and SAR time series and the Time-Series-to-Vision representation can reduce some of these errors by integrating complementary phenological and backscatter information, local confusion in complex landscapes cannot be completely eliminated. We have therefore added a corresponding discussion to Discussion Section 5.4 and noted that future work could further improve the accuracy and robustness of GlobalRice20 by incorporating more detailed ancillary data, remote-sensing observations with higher spatial and temporal resolution, and additional reference samples from areas prone to confusion.

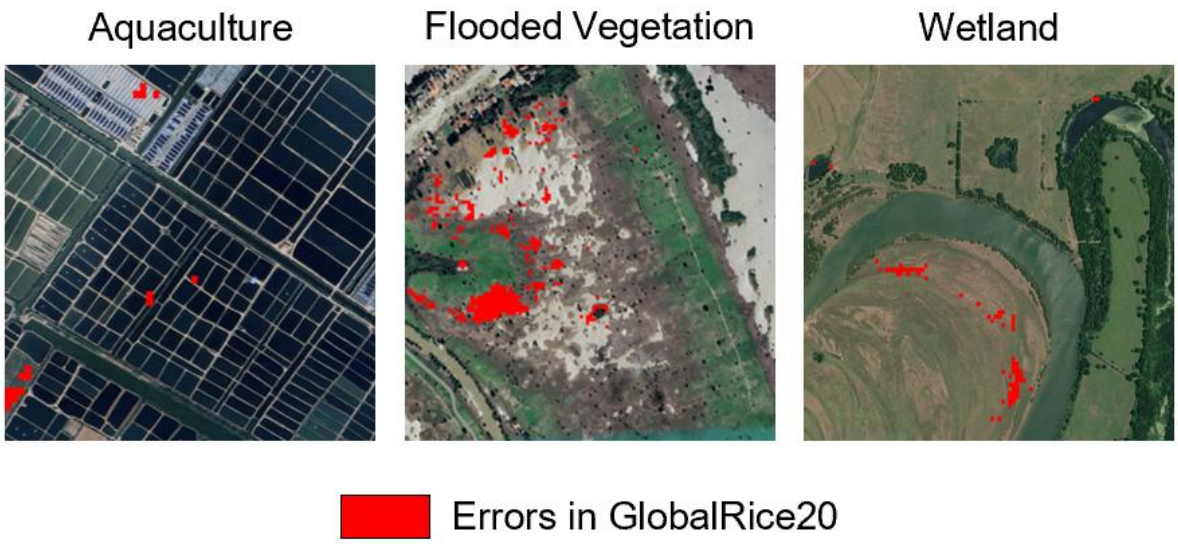


**Figure R2. Representative examples of potential confusion between mapped paddy rice and inundated land-cover types.**

We sincerely thank the reviewer again for this constructive comment. The corresponding analyses and discussion have been added to the Discussion section of the revised manuscript. The specific revisions are provided below.

**Pages 28-29**

**5.1 Effects of Interannual Data Configurations on Classification Performance**

To compare model performance under the different optical sensor and SAR data configurations in 2015 and 2024, we conducted modality ablation experiments separately using data from each year. Landsat-8 and Sentinel-1 were used for 2015, while Sentinel-2 and Sentinel-1 were used for 2024, with three input configurations for each year: Optical-only, SAR-only, and Optical+SAR. The same T2VRCM network architecture and evaluation metrics were used for both years, and the models were trained and tested separately using data from the corresponding year, allowing classification performance to be compared under the optical and SAR data configurations of different years.

The experimental results are shown in Table 3. The combined Optical+SAR input achieved the highest classification accuracy in both years. In 2015, the F1 values for Optical+SAR, Optical-only, and SAR-only were 90.24%, 85.46%, and 81.05%,

respectively, while the corresponding values in 2024 were 92.32%, 87.19%, and 86.17%, indicating clear complementarity between optical and SAR information in both years. Specifically, in 2015, Optical+SAR improved F1 over Optical-only and SAR-only by 4.78 and 9.19 percentage points, respectively; in 2024, the corresponding improvements were 5.13 and 6.15 percentage points, respectively.

**Table 3. Classification performance of T2VRCM under different modality configurations in 2015 and 2024.**

| Year | Setting | UA (%) | PA (%) | F1 (%) | OA (%) |
|---|---|---|---|---|---|
| 2015 | Full (L8 + S1) | 90.11 | 90.37 | 90.24 | 90.23 |
| 2015 | Optical-only (L8) | 84.93 | 86.00 | 85.46 | 85.37 |
| 2015 | SAR-only (S1) | 81.19 | 80.91 | 81.05 | 81.08 |
| 2024 | Full (S2 + S1) | 92.52 | 92.12 | 92.32 | 92.33 |
| 2024 | Optical-only (S2) | 86.76 | 87.62 | 87.19 | 87.13 |
| 2024 | SAR-only (S1) | 85.80 | 86.54 | 86.17 | 86.11 |

In terms of interannual differences, Optical-only F1 was 1.73 percentage points higher in 2024 than in 2015. This difference is mainly associated with the different optical data sources and observation conditions in the two years. Compared with Landsat-8, Sentinel-2 has certain advantages in temporal resolution and the acquisition of valid observations, which support a more complete characterization of temporal variations throughout the rice growing cycle and thereby improve the ability to distinguish rice from non-rice. Meanwhile, SAR-only F1 was 5.12 percentage points higher in 2024 than in 2015, indicating that differences in Sentinel-1 data availability and coverage conditions between years also have an important effect on rice classification performance. Under the full multimodal configuration, F1 was 2.08 percentage points higher in 2024 than in 2015. Overall, the fusion of optical and SAR data produced consistent performance gains in both years, while the classification differences between 2015 and 2024 reflect the effects of sensor configurations and data

coverage conditions in different years on the classification performance of GlobalRice20.

**Pages 29-30**

### 5.2 Effects of Sentinel-1 Data Coverage on Classification Performance

To further analyze the effects of Sentinel-1 data coverage conditions on classification performance, we conducted a data coverage sensitivity analysis using the independent test samples from 2024. Based on the number of valid Sentinel-1 12 d composite time steps during 2024, the test samples were divided into three coverage categories: No coverage, Partial coverage, and Complete coverage. No coverage indicates that no valid Sentinel-1 observations were available during 2024; Partial coverage indicates that some valid Sentinel-1 time steps were available, but the annual time series was incomplete; and Complete coverage indicates that valid Sentinel-1 observations were available for all 31 of the 12 d compositing periods. Keeping the models, test-sample predictions, and evaluation methods unchanged, we calculated F1 separately for RF, LSTM, Transformer, and each visual classification model under the three coverage conditions.

The experimental results are shown in Table 4. All models exhibited a consistent pattern of performance variation: F1 increased progressively with improving Sentinel-1 data coverage, and all models achieved their highest accuracy under Complete coverage and their lowest performance under No coverage. This indicates that the dynamic backscatter information contained in Sentinel-1 time series can effectively complement optical time-series features. T2VRCM achieved F1 values of 86.56%, 92.16%, and 93.06% under No coverage, Partial coverage, and Complete coverage, respectively, attaining the highest classification accuracy in all three coverage categories.

**Table 4. F1 of different models under different Sentinel-1 coverage conditions.**

| Model | No coverage F1 (%) | Partial coverage F1 (%) | Complete coverage F1 (%) |
|---|---|---|---|
| RF | 75.11 | 82.09 | 84.47 |
| LSTM | 76.83 | 84.10 | 86.33 |
| Transformer | 77.73 | 84.86 | 86.84 |
| ResNet-18 | 80.91 | 87.30 | 88.87 |
| ViT-Tiny | 81.93 | 87.78 | 89.39 |
| Swin-Tiny | 82.02 | 88.25 | 89.62 |
| ConvNeXt-Atto | 82.66 | 88.36 | 89.67 |
| MambaOut-Femto | 83.08 | 89.14 | 90.22 |
| T2VRCM (Ours) | 86.56 | 92.16 | 93.06 |

Further analysis of T2VRCM performance across the different coverage categories shows that F1 differed by only 0.90 percentage points between Partial coverage and Complete coverage. This indicates that, when the annual SAR time series retains some valid observations, partial temporal missingness has a relatively limited effect on classification performance. In contrast, when Sentinel-1 data were completely missing, F1 decreased from 93.06% under Complete coverage to 86.56%, a decline of 6.50 percentage points. Because information from the entire SAR modality is unavailable in this case, the model cannot obtain the annual backscatter dynamics provided by Sentinel-1, and the complementary information between optical and SAR data is also lost, resulting in a more pronounced decline in classification performance. Compared with the other models, T2VRCM exhibited the smallest performance declines when moving from Complete coverage to either Partial coverage or No coverage, indicating good stability under changes in Sentinel-1 data coverage. Overall, missing some SAR time steps has a relatively limited effect on T2VRCM classification performance, whereas complete SAR modality absence causes more pronounced information loss and

a greater decline in accuracy. This indicates that Sentinel-1 data availability and the completeness of annual coverage are important factors affecting the classification performance of GlobalRice20.

**Pages 31-32**

### 5.4 Limitations and Future Work

Although GlobalRice20 demonstrates strong overall performance in terms of mapping accuracy and spatial coverage, several limitations remain and warrant further investigation.

First, the mapping accuracy is partly constrained by the spatial resolution of the available optical satellite observations. This limitation is more evident for 2015, when Landsat-8 data were used, and likely contributed to the relatively lower mapping accuracy compared with 2024. With the continued improvement of satellite observations in both spatial and temporal resolution, the delineation of rice field boundaries and the representation of fine-scale spatial patterns are expected to be further improved.

Second, the current study focuses on mapping the spatial distribution of rice at the global scale and does not distinguish among different rice cropping systems, such as single- and double-cropping rice. This limits the direct use of GlobalRice20 for characterizing cropping intensity and production dynamics. Future work will therefore extend the Time-Series-to-Vision framework toward global rice cropping-intensity mapping, thereby further enhancing the value of GlobalRice20 for food production assessment and SDG 2 monitoring.

Third, for the native time-series baseline models, linear interpolation was adopted to regularize incomplete temporal observations. This strategy provides a stable and computationally efficient solution for large-scale experiments, but it may not fully capture complex patterns of missing observations. More advanced temporal interpolation and missing-data handling approaches may further improve the ability of time-series models to process irregular and incomplete observations. Evaluating the

effectiveness and scalability of these approaches in global crop-mapping applications therefore represents an important direction for future research.

Finally, GlobalRice20 still contains uncertainties arising from both class confusion and regional heterogeneity. In particular, rice fields may be confused with land-cover types exhibiting similar hydrological or phenological characteristics, including aquaculture areas, wetlands, and flooded vegetation. Additional uncertainties may also arise from regional differences in cropping systems, field management practices, and image quality. Although the fusion of multi-source time-series observations helps reduce some of these errors, local misclassification cannot be completely avoided. Future studies could incorporate more detailed ancillary data, observations with higher spatial and temporal resolution, and additional reference samples to further improve the accuracy and robustness of GlobalRice20 in complex landscapes.

**7. The time-series-to-vision transformation discards numerical precision. Encoding floating-point time-series values as colored lines on a 224×224 pixel image introduces quantization error. The model must learn to "read" a graph visually rather than operating on the underlying numbers. While the results suggest this works in practice, the paper should acknowledge and discuss this trade-off. A simple experiment comparing the T2VRCM backbone fed with raw numerical input (appropriately padded/masked) versus the image-based input would clarify how much accuracy is gained or lost by the visual encoding itself.**

**RESPONSE:** Thank you very much for this valuable comment. In response to your suggestion to further quantify the potential gain or loss associated with the visual encoding itself, we conducted an additional controlled representation-level comparison using the 2024 dataset to further distinguish the contribution of the T2V visual representation from that of the subsequent network architecture. It should be noted that the Stem, InceptionDWC, and FSFI modules in T2VRCM are specifically designed for two-dimensional image features. Therefore, one-dimensional numerical time series

cannot be directly fed into the identical T2VRCM architecture without modifying its input structure. To minimize differences in network architecture, we adopted the same Transformer encoder and classification head for both settings, changing only the input representation and the corresponding input embedding.

Specifically, the first experiment used numerical time-series representations of EVI, LSWI, VV, and VH and followed the preprocessing procedure and Transformer baseline configuration used in Table 1. In the second experiment, the same multi-source time series were converted into T2V images and fed into an identically configured Transformer encoder through image patch embedding. Both experiments were conducted using the 2024 dataset with exactly the same training and test sets. The Transformer depth, hidden dimension, number of attention heads, classification head, and training parameters were kept identical. Because numerical time-series representations and two-dimensional images have inherently different data structures, each setting used an embedding module appropriate for its input format, while the subsequent Transformer encoder and classification head remained unchanged, thereby minimizing the influence of architectural differences on the comparison.

The experimental results are shown in Table R3. When using the numerical time-series representation, the Transformer achieved UA, PA, F1, and OA values of 85.87%, 85.03%, 85.45%, and 85.52%, respectively. After adopting the T2V image representation, the corresponding metrics increased to 88.05%, 87.77%, 87.91%, and 87.93%, respectively. In particular, F1 and OA improved by 2.46 and 2.41 percentage points, respectively. These results demonstrate that, under the same core Transformer encoder and training settings, the T2V image representation achieves higher classification accuracy and provides more direct experimental evidence for evaluating the contribution of the T2V representation itself.

**Table R3. Comparison of numerical time-series and T2V image representations using the same Transformer encoder.**

| Representation | Model | UA (%) | PA (%) | F1 (%) | OA (%) |
|---|---|---|---|---|---|
| Numerical time series | Transformer | 85.87 | 85.03 | 85.45 | 85.52 |
| T2V image | Transformer | 88.05 | 87.77 | 87.91 | 87.93 |

The results further show that, although converting continuous floating-point time series into 224 × 224 pixel line-chart images inevitably introduces some loss of numerical precision due to image resolution, line width, and pixel discretization, the T2V representation still achieved higher classification accuracy under the same Transformer encoder. This indicates that, under the task and experimental settings considered in this study, the loss of numerical precision caused by pixel discretization does not offset the overall classification gain provided by the T2V representation. This improvement mainly arises because T2V reorganizes irregular multi-source time series into a two-dimensional visual representation that jointly captures temporal positions, overall curve shapes, stage-wise variations, and relationships among different remote-sensing variables. This enables the model to more effectively extract structural characteristics associated with rice phenological development, rather than relying primarily on the precise pointwise preservation of individual temporal values.

In addition, this newly added experiment complements the network-module ablation study reported in Table 2. The present experiment mainly quantifies the effect of the input representation on classification performance, whereas Table 2 further evaluates the contributions of network components such as InceptionDWC and FSFI. Together, these two analyses provide a clearer distinction between the contribution of the T2V visual encoding strategy and that of the T2VRCM network design to the final performance improvement. We sincerely thank the reviewer again for this valuable suggestion.

The specific revisions are provided below.

**Pages 30-31**

### 5.3 Effects of the Time-Series-to-Vision Representation on Classification Performance

To further distinguish the contribution of the T2V visual representation from that of the subsequent network architecture, we conducted a controlled representation-level comparison based on the 2024 dataset. Because the Stem, InceptionDWC, and FSFI modules in T2VRCM are specifically designed for two-dimensional image features, one-dimensional numerical time series cannot be directly fed into the identical T2VRCM architecture without modifying the input structure. Therefore, the same Transformer encoder and classification head were adopted for comparison. The first setting used numerical time-series representations of EVI, LSWI, VV, and VH, following the preprocessing procedure and Transformer baseline configuration used in Table 1. The second setting converted the same multi-source time series into T2V images and fed them into the same Transformer encoder through image patch embedding. The two experiments used exactly the same training and test sets and maintained identical Transformer depth, hidden dimension, number of attention heads, classification head, and training parameters, thereby minimizing the influence of differences in network architecture and training settings on the comparison.

The experimental results are shown in Table 5. Using the numerical time-series representation, the Transformer achieved UA, PA, F1, and OA values of 85.87%, 85.03%, 85.45%, and 85.52%, respectively. After adopting the T2V image representation, the corresponding metrics increased to 88.05%, 87.77%, 87.91%, and 87.93%. Under the same core Transformer encoder and major training settings, the T2V representation improved F1 and OA by 2.46 and 2.41 percentage points, respectively, demonstrating that transforming multi-source time series into a two-dimensional visual representation can further improve rice classification performance.

**Table 5. Comparison of numerical time-series and T2V image representations using the same Transformer encoder.**

| Representation | Model | UA (%) | PA (%) | F1 (%) | OA (%) |
|---|---|---|---|---|---|
| Numerical time series | Transformer | 85.87 | 85.03 | 85.45 | 85.52 |
| T2V image | Transformer | 88.05 | 87.77 | 87.91 | 87.93 |

The performance improvement mainly stems from the way T2V reorganizes irregular multi-source time-series structures. Specifically, T2V maps the actual temporal positions of different remote-sensing variables, the overall annual curve shapes, stage-wise variations, and the relationships among EVI, LSWI, VV, and VH into a unified two-dimensional visual space, thereby transforming information originally distributed across different time steps and variables into an integrated representation with explicit structural relationships. Based on this representation, the model can simultaneously exploit temporal positions, multivariable variation relationships, and overall phenological curve characteristics for classification, enabling more effective extraction of structural information associated with rice growth processes. Although converting continuous numerical time series into images inevitably introduces some loss of numerical precision, the experimental results show that this loss does not offset the advantages provided by the structured visual representation, ultimately resulting in an overall improvement in classification accuracy.

**8. Reference sample construction lacks detail. The phrase "rigorous visual interpretation process integrating high-resolution optical imagery and existing regional rice maps" is vague. How many interpreters were involved? What was the interpretation protocol? Was inter-annotator agreement measured? Which high-resolution imagery was used, and at what date? These details are important for assessing label quality.**

**RESPONSE:** Thank you for your valuable feedback. We fully agree that clearly detailing the composition of the annotation team, the interpretation protocol, and the specific sources and timing of the imagery used is crucial for readers to evaluate the quality of the labels. In response to your four specific questions, our detailed replies are as follows, and the relevant content has been incorporated into Section 2.3.1 of the revised manuscript.

(1) Regarding the number and composition of the annotators

The annotation of the reference samples was collaboratively completed by a dedicated annotation team. The members of this team included: 8 Master's and Ph.D. students, 4 professors with crop mapping experience in the field of agricultural remote sensing, and 1 invited expert from the Academy of Agricultural Sciences. During the sample generation process, the expert from the Academy of Agricultural Sciences provided specific guidance on the visual interpretation rules for identifying paddy rice fields in optical imagery. Additionally, the team held multiple meetings to discuss matters related to sample annotation. Therefore, the sample annotation in this study was accomplished through close collaboration by a multi-tiered team of experienced experts and scholars.

(2) Regarding the interpretation protocol

The annotators selected spatially discrete and mutually independent candidate locations within the rice-cultivating regions of each country. They conducted visual interpretation referencing the annual composite optical imagery corresponding to the target years, as well as existing regional rice maps. Integrated with the "stable sample" strategy, only candidate points that exhibited a consistent land cover class (either persistent rice or persistent non-rice) in both the 2015 and 2024 observations were ultimately included in the reference sample set. This protocol has been explicitly clarified in the revised manuscript across three aspects: spatial site selection, interpretation basis, and temporal consistency constraints.

(3) Regarding the measurement of inter-annotator agreement

We understand your desire to see a quantitative assurance of label quality. It should be noted that this study did not calculate statistical metrics for inter-annotator agreement. Instead, we adopted a more rigorous strategy that combined team discussions with field investigations: the annotation of all samples was conducted under unified rules guided by the expert. The team not only verified and confirmed the correctness of the annotations utilizing their extensive theoretical knowledge and experience but also reached conclusions through team consensus meetings whenever ambiguities were encountered. This collaboration and review mechanism fundamentally eliminated potential discrepancies among annotators, providing an assurance of label consistency and accuracy that is no less robust than statistical metrics derived from sampling. We have explicitly described this quality control mechanism in the revised manuscript so that readers can evaluate the label quality accordingly.

(4) Regarding the imagery used and its specific timing

The imagery relied upon for annotation consisted of optical imagery corresponding to the target years, specifically including high-resolution Google Earth/Maps imagery, as well as Landsat-8 imagery for 2015 and Sentinel-2 imagery for 2024. Additionally, we incorporated existing regional rice maps as auxiliary references during the interpretation. The term "high-resolution optical imagery" mentioned in the original manuscript primarily referred to the high-resolution Google Earth imagery used to assist in the interpretation. To avoid any misunderstanding, we have refined this expression in the revised manuscript, explicitly listing the specific data sources used and their corresponding years.

In summary, we have supplemented Section 2.3.1 of the revised manuscript with detailed explanations regarding the composition of the annotation team, the two-stage quality control process, the label consistency assurance mechanism, and the specific imagery sources with their corresponding years. We have also corrected the imprecise expressions from the original manuscript to enhance the transparency and rigor of the sample construction process. Thank you once again for your constructive comments, which have significantly helped us refine the descriptions in the methodology section.

The specific revisions are as follows:

**Page 6-7**

All samples were derived through rigorous visual interpretation of discrete locations distributed across rice-cultivating regions in each country, rather than being sampled from continuous spatial blocks. Consequently, this design mitigates the likelihood of adjacent or duplicate samples appearing in both the training and test sets, thereby reducing the risk of accuracy overestimation driven by spatial autocorrelation.

The sample annotation primarily relied on high-resolution optical imagery corresponding to the target years (2015 and 2024), including Google Earth, Landsat-8, and Sentinel-2 imagery. Additionally, existing regional rice maps were incorporated to locate potential cultivation extents. The annotation process was collaboratively executed by a dedicated team comprising 8 graduate students, 4 professors with extensive experience in agricultural remote sensing, and 1 expert from the Academy of Agricultural Sciences. This process involved multiple stages: initially, the expert provided unified guidance and established standardized visual interpretation rules for the team; subsequently, the 8 graduate students conducted preliminary labeling of candidate samples based on these criteria; finally, the 4 professors comprehensively audited the preliminary results. Any discrepancies encountered during the review were resolved through multiple team consensus meetings, thereby ensuring the high accuracy and reliability of the final labels.

**9. Cropping system not distinguished in the output. The paper discusses single-season, double-season, and mixed-season rice systems (Section 2.1) but the final product is a binary rice/non-rice map. For a 20 m product aimed at SDG 2 monitoring, distinguishing cropping intensity would add substantial value. This should at least be discussed as a limitation or future direction.**

**RESPONSE:** Thank you very much for your valuable suggestion. We agree that further

distinguishing between single-cropping rice, double-cropping rice, and mixed cropping systems could substantially enhance the application value of GlobalRice20 for agricultural production capacity assessment and SDG 2 monitoring. However, the primary objective of the present study is to develop a globally consistent annual rice distribution product. Therefore, the task is defined as a binary classification of rice and non-rice areas. The description of single-cropping, double-cropping, and mixed cropping systems in Section 2.1 is primarily intended to illustrate the substantial heterogeneity in phenology and cropping systems across global rice-growing regions.

The current GlobalRice20 product does not further distinguish whether rice is cultivated once or multiple times within the same pixel during a year. We acknowledge that mapping global rice cropping intensity and identifying different cropping systems, such as single- and double-cropping systems, represent important directions for future research. In future work, we plan to further leverage the Time-Series-to-Vision framework to model multi-cycle phenological information, thereby enhancing the value of the dataset for assessing rice cropping intensity, production potential, and monitoring progress toward SDG 2. We have added a corresponding discussion in the Discussion section of the revised manuscript. Once again, we sincerely appreciate your valuable suggestion.

The specific revisions are as follows.

**Page 32**

Second, the current study focuses on mapping the spatial distribution of rice at the global scale and does not distinguish among different rice cropping systems, such as single- and double-cropping rice. This limits the direct use of GlobalRice20 for characterizing cropping intensity and production dynamics. Future work will therefore extend the Time-Series-to-Vision framework toward global rice cropping-intensity mapping, thereby further enhancing the value of GlobalRice20 for food production assessment and SDG 2 monitoring.

**10. Computational cost is not reported. Global-scale inference at 20 m resolution is a massive undertaking. The paper does not report total processing time, number of GEE computation hours, or the cost of running T2VRCM inference globally. This information is relevant for reproducibility and for other groups wishing to update or extend the dataset.**

**RESPONSE:** Thank you very much for this valuable suggestion. We agree that global rice mapping at a 20 m resolution involves large-scale processing of multi-source remote sensing data and extensive model inference. Reporting the actual processing time and computational configuration is important for assessing the reproducibility of GlobalRice20 and for providing a practical reference for other research groups seeking to update or extend the dataset. The production of GlobalRice20 mainly consists of two stages. First, Google Earth Engine (GEE) was used for global multi-source remote sensing data preprocessing, tiled export, and local download. Subsequently, T2VRCM inference was performed locally. To improve inference efficiency, the global inference task was executed in parallel on four independent workstations, each equipped with one NVIDIA RTX 3090 GPU.

For the 2015 and 2024 global datasets, GEE preprocessing, export, and local download required approximately 52 days of wall-clock time, while T2VRCM inference required approximately 85 days. If these two stages had been executed entirely sequentially, the cumulative wall-clock time would have been approximately 137 days. In the actual production workflow, however, T2VRCM inference for individual regions was initiated as soon as the corresponding data had been exported and downloaded, thereby forming a pipelined workflow in which data processing and model inference partially overlapped. As a result, the actual overall wall-clock processing time for producing the two GlobalRice20 products was reduced to approximately 107 days. The detailed processing times are summarized below.

**Table R4. Processing time for GlobalRice20 production.**

| Processing stage | Wall-clock time for 2015 + 2024 |
|---|---|
| GEE preprocessing, export, and local download | 52 days |
| T2VRCM inference | 85 days |
| Sequential processing | 137 days |
| Actual GlobalRice20 production | 107 days |

Following your suggestion, we have added the computational configuration, major processing stages, and actual processing time to the revised manuscript. This additional information improves the transparency of the GlobalRice20 production workflow and provides a practical reference regarding computational resources and processing time for other research groups seeking to update or extend the dataset. We sincerely thank the reviewer again for this valuable suggestion. The specific revision is provided below.

**Page 20**

The production of GlobalRice20 also involved substantial computational processing at the global scale. Using four workstations, each equipped with an NVIDIA RTX 3090 GPU, GEE data processing and local export for the 2015 and 2024 products required approximately 52 days, while local T2VRCM inference required approximately 85 days. Because the two processes were conducted in parallel, the actual total production time for the two GlobalRice20 products was approximately 107 days.

**11. SAR data compositing at 12-day intervals is stated but not justified. Why 12 days? Sentinel-1A has a 12-day repeat cycle, but in regions with ascending and descending passes, the effective revisit is 6 days. Was orbit direction handled? Were ascending and descending passes composited together or separately? Mixing orbit directions without accounting for viewing geometry differences can introduce noise in backscatter values.**

**RESPONSE:** We thank the reviewer for pointing out this important detail regarding Sentinel-1 data processing. We used a 12-day compositing interval mainly because the SAR data for both target years in this study were primarily acquired from Sentinel-1A. Given that Sentinel-1A has a nominal repeat cycle of 12 days, using a consistent 12-day temporal window helps ensure comparability in the SAR time-series representation between 2015 and 2024. In addition, we applied local incidence angle normalization to the Sentinel-1 backscatter data to reduce variations caused by differences in acquisition geometry among observations. Previous studies have shown that incidence angle normalization facilitates the combined use of Sentinel-1 observations acquired at different incidence angles and along different orbits, and also helps reduce orbit-related differences in backscatter values (Kaplan et al., 2021; Arias et al., 2022). We have clarified this point in the revised manuscript. Once again, we sincerely appreciate your valuable suggestion.

The specific revisions are as follows:

**Page 4-5**

Both VV and VH polarizations were utilized, and ascending and descending acquisitions were aggregated together into 12-day mean composites after local incidence-angle normalization to reduce viewing-geometry-related backscatter differences (Kaplan et al., 2021; Arias et al., 2022).

**References**

Arias, M., Campo-Bescós, M. Á., and Álvarez-Mozos, J.: On the influence of acquisition geometry in backscatter time series over wheat, International Journal of Applied Earth Observation and Geoinformation, 106, 102671, 2022.

Kaplan, G., Fine, L., Lukyanov, V., Manivasagam, V. S., Tanny, J., and Rozenstein, O.: Normalizing the Local Incidence Angle in Sentinel-1 Imagery to Improve Leaf Area Index, Vegetation Height, and Crop Coefficient Estimations, Land, 10, 680, 2021.

**12. The ablation study is limited in scope. The ablation in Table 2 tests InceptionDWC and FSFI independently but does not ablate the time-series-to-vision transformation itself. A comparison of the T2VRCM backbone operating on raw time-series input (maybe with proper missing-data handling) versus image input would isolate the contribution of the visual encoding strategy from the contribution of the network architecture.**

**RESPONSE:** Thank you very much for this valuable suggestion. We agree that the ablation study in the original manuscript mainly focused on network components such as InceptionDWC and FSFI, which demonstrated the effectiveness of the internal architecture of T2VRCM, but did not independently quantify the contribution of the Time-Series-to-Vision (T2V) representation itself to classification performance. Following your suggestion, we conducted an additional controlled representation-level comparison using the 2024 dataset to further distinguish the contribution of the T2V visual representation from that of the subsequent network architecture.

It should be noted that the Stem, InceptionDWC, and FSFI modules in T2VRCM are specifically designed for two-dimensional image features. Therefore, one-dimensional numerical time series cannot be directly fed into the identical T2VRCM architecture without modifying the input structure. To minimize differences in network architecture, we adopted the same Transformer encoder and classification head for comparison, changing only the input representation and the corresponding input embedding. In the first setting, EVI, LSWI, VV, and VH were represented as numerical time series, following the Transformer configuration used in Table 1. In the second setting, the same multi-source time series were converted into T2V images and fed into an identically configured Transformer encoder through image patch embedding. Both experiments were conducted using the 2024 dataset with exactly the same training and test sets, while the Transformer depth, hidden dimension, number of attention heads, classification head, and training parameters were kept identical. Because numerical time-series representations and two-dimensional images have inherently different data structures, each setting used an embedding module appropriate for its input format,

whereas the subsequent Transformer encoder and classification head remained unchanged. This design minimizes the influence of architectural differences and allows the comparison to focus primarily on the effect of the input representation.

The experimental results are shown in Table R5. Using the numerical time-series representation, the Transformer achieved UA, PA, F1, and OA values of 85.87%, 85.03%, 85.45%, and 85.52%, respectively. After adopting the T2V image representation, the corresponding metrics increased to 88.05%, 87.77%, 87.91%, and 87.93%, respectively. In particular, F1 and OA improved by 2.46 and 2.41 percentage points, respectively. These results show that, under the same core Transformer encoder and training settings, the T2V image representation achieves higher classification accuracy and provides more direct experimental evidence for evaluating the contribution of the T2V representation itself.

**Table R5. Ablation of numerical time-series and T2V image representations using the same Transformer encoder.**

| Representation | Model | UA (%) | PA (%) | F1 (%) | OA (%) |
|---|---|---|---|---|---|
| Numerical time series | Transformer | 85.87 | 85.03 | 85.45 | 85.52 |
| T2V image | Transformer | 88.05 | 87.77 | 87.91 | 87.93 |

Based on this additional experiment, the revised ablation analysis now consists of two complementary levels. The newly added experiment is intended to quantify the contribution of the input representation to classification performance. The results show that transforming irregular multi-source time series into a two-dimensional visual representation capable of expressing temporal positions, curve shapes, stage-wise variations, and relationships among different remote-sensing variables improves classification performance in the current task. The original Table 2, in contrast, further evaluates the additional gains provided by the network architecture specifically designed for T2V images through the ablation of InceptionDWC and FSFI. Therefore, these two sets of experiments evaluate the proposed method from the complementary

perspectives of input representation and network architecture, respectively, providing a clearer distinction between the contributions of the T2V visual representation strategy and the T2VRCM network design to the final performance improvement and further strengthening the ablation analysis of the manuscript. The corresponding experimental results and discussion have been added to Discussion Section 5.3 of the revised manuscript. We sincerely thank the reviewer again for this valuable suggestion.

The specific revisions are provided below.

**Pages 30-31**

**5.3 Effects of the Time-Series-to-Vision Representation on Classification Performance**

To further distinguish the contribution of the T2V visual representation from that of the subsequent network architecture, we conducted a controlled representation-level comparison based on the 2024 dataset. Because the Stem, InceptionDWC, and FSFI modules in T2VRCM are specifically designed for two-dimensional image features, one-dimensional numerical time series cannot be directly fed into the identical T2VRCM architecture without modifying the input structure. Therefore, the same Transformer encoder and classification head were adopted for comparison. The first setting used numerical time-series representations of EVI, LSWI, VV, and VH, following the preprocessing procedure and Transformer baseline configuration used in Table 1. The second setting converted the same multi-source time series into T2V images and fed them into the same Transformer encoder through image patch embedding. The two experiments used exactly the same training and test sets and maintained identical Transformer depth, hidden dimension, number of attention heads, classification head, and training parameters, thereby minimizing the influence of differences in network architecture and training settings on the comparison.

The experimental results are shown in Table 5. Using the numerical time-series representation, the Transformer achieved UA, PA, F1, and OA values of 85.87%, 85.03%, 85.45%, and 85.52%, respectively. After adopting the T2V image

representation, the corresponding metrics increased to 88.05%, 87.77%, 87.91%, and 87.93%. Under the same core Transformer encoder and major training settings, the T2V representation improved F1 and OA by 2.46 and 2.41 percentage points, respectively, demonstrating that transforming multi-source time series into a two-dimensional visual representation can further improve rice classification performance.

**Table 5. Comparison of numerical time-series and T2V image representations using the same Transformer encoder.**

| Representation | Model | UA (%) | PA (%) | F1 (%) | OA (%) |
|---|---|---|---|---|---|
| Numerical time series | Transformer | 85.87 | 85.03 | 85.45 | 85.52 |
| T2V image | Transformer | 88.05 | 87.77 | 87.91 | 87.93 |

The performance improvement mainly stems from the way T2V reorganizes irregular multi-source time-series structures. Specifically, T2V maps the actual temporal positions of different remote-sensing variables, the overall annual curve shapes, stage-wise variations, and the relationships among EVI, LSWI, VV, and VH into a unified two-dimensional visual space, thereby transforming information originally distributed across different time steps and variables into an integrated representation with explicit structural relationships. Based on this representation, the model can simultaneously exploit temporal positions, multivariable variation relationships, and overall phenological curve characteristics for classification, enabling more effective extraction of structural information associated with rice growth processes. Although converting continuous numerical time series into images inevitably introduces some loss of numerical precision, the experimental results show that this loss does not offset the advantages provided by the structured visual representation, ultimately resulting in an overall improvement in classification accuracy.

**13. Some figures are bit of vague. I would recommend authors to use vector graph instead of raster graph.**

**RESPONSE:** Thank you very much for this valuable suggestion. Following your comment, we carefully re-examined and optimized the figures throughout the manuscript to improve the clarity of the text, lines, and overall visual presentation. The corresponding revisions have been completed in the revised manuscript. We sincerely appreciate your helpful suggestion.

**14. Duplicate reference. Ni et al. (2021) appears twice in the reference list (lines 581–586), with identical content.**

**RESPONSE:** We sincerely apologize for the duplicated reference entry. Thank you for pointing out this oversight. We have carefully checked the reference list and removed the duplicate Ni et al. (2021) citation. The reference list has now been corrected in the revised manuscript. Once again, we sincerely appreciate your valuable suggestion.

# Response to Reviewer 2

**Comments to the Author**

**GlobalRice20 gives us the first validated, globally consistent 20 m paddy rice map for 2015 and 2024, built from multi-source optical and SAR data through the Time-Series-to-Vision framework. It is a timely and genuinely useful contribution to food-security and SDG 2 work. However, before finalizing the published paper, I believe that some of its content still requires further clarification or modification.**

**RESPONSE:** We sincerely appreciate your positive assessment of GlobalRice20 and your thorough and insightful review of the manuscript. Your detailed comments and suggestions have been very helpful in strengthening the methodological description, scientific interpretation, and overall quality of the paper. We have carefully considered each comment and provided corresponding point-by-point responses below.

**1. Validation design and accuracy estimation**

**Validation is the most problematic component of the current manuscript. The authors report using approximately 164,000 reference samples, but the sampling design is not adequately described. Was the sample generated using stratified random sampling, simple random sampling, systematic sampling, or another design? The sampling procedure, strata, inclusion probabilities, and allocation across regions and classes should be clearly documented.**

**The authors also adopted an approximately 1:1 ratio of rice to non-rice samples. Because this sample composition does not reflect the actual proportions of rice and non-rice areas in the mapped population, the reported overall accuracy cannot be interpreted as a design-unbiased estimate of map accuracy unless appropriate area-based weighting is applied. Class proportions should be adjusted according to mapped area, following the good-practice recommendations of Olofsson et al. (2014). As currently presented, the overall accuracy appears to have been**

**calculated directly from the balanced sample and may therefore be substantially biased. The authors should recalculate the accuracy metrics using the appropriate sampling weights and report uncertainty estimates, including confidence intervals.**

**Reference:**

**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, 42–57.**

**RESPONSE:** Thank you for your professional and constructive comments regarding the validation design and accuracy estimation. Your feedback consists of two parts, which we address separately below.

(1) Clarification of the Sampling Design

We fully agree that clearly articulating the sampling design of the reference samples is crucial for the transparency and rigor of the label generation process. In this study, a total of 164,000 stable samples were constructed (131,200 for training and 32,800 for testing), corresponding to a training-to-test ratio of 4:1. We employed a spatially stratified random sampling strategy, utilizing each country as a spatial stratification unit. Within each country, samples were independently and randomly partitioned into the training and test sets, strictly maintaining the 4:1 ratio. Simultaneously, we ensured that the 1:1 balance between rice and non-rice samples within each country was preserved in both the training and test sets. This country-based stratification guarantees that the model is adequately and evenly trained and validated across different climate zones and cropping systems.

To further evaluate the model's performance under dynamic land cover conditions, we constructed an additional, independent dynamic test set. Candidate locations were first identified based on the differences between the 2015 and 2024 GlobalRice20 maps. Subsequently, spatially stratified random sampling was employed to extract 16,400 points for each target year, ensuring a balanced representation between two change trajectories: newly established paddies and abandoned paddies. All candidate points

were collaboratively labeled by a dedicated annotation team comprising 8 graduate students, 4 professors with agricultural remote sensing expertise, and 1 expert from the Academy of Agricultural Sciences. Guided by standardized visual interpretation rules established by the expert, the 8 graduate students independently determined the true land cover classes for 2015 and 2024 using the corresponding optical imagery. Subsequently, the 4 professors comprehensively audited the annotation results, resolving any ambiguities through field investigations and team consensus meetings. This process was entirely independent of the model's own predictions. The labeling criteria remained strictly consistent with the original 164,000 stable samples.

The intra-stratum sampling was conducted via equal-probability random selection. All candidate points were spatially discrete and mutually independent locations, rather than being drawn from continuous spatial blocks, thereby mitigating the risk of accuracy overestimation driven by spatial autocorrelation.

(2) Area-Weighted Accuracy Estimation

We entirely agree with your assessment that because the rice/non-rice samples were collected at a 1:1 ratio, they do not reflect the true area proportion of paddy rice within the overall mapped extent. Therefore, the overall accuracy calculated directly from these samples in the original manuscript can only be considered sample-level accuracy and cannot be directly interpreted as an unbiased estimate of the map accuracy. To address this issue, we strictly followed the good practice guidelines proposed by Olofsson et al. (2014). Using the obtained map area statistics and the confusion matrix of the independent test set, we recalculated the area-weighted accuracy metrics and reported the uncertainty estimates. The specific procedures are as follows.

We utilized the two classes of GlobalRice20, "rice" and "non-rice", as stratification variables, constrained by the GADM data of the 98 countries, to calculate the area weight for each stratum:

$$W_h = \frac{A_h}{A_{tot}}, \mathrm{h} \in \{\mathrm{R}, \mathrm{N}\}$$

Here, $A_h$ represents the mapped area of stratum $h$ (the mapped rice stratum $R$ or mapped non-rice stratum $N$), and $A_{tot} = A_R + A_N$ represents the total area of the study region for both classes combined. We would like to clarify a conceptual distinction regarding the area used here: the rice area used to calculate the stratum weights is the planted area derived from pixel-by-pixel statistics of GlobalRice20, which differs from the sown area of rice reported in the main text. The calculation results are shown in the table below:

| **Year** | $\mathbf{W_R}$ | $\mathbf{W_N}$ |
|---|---|---|
| 2015 | 5.86% | 94.14% |
| 2024 | 6.26% | 93.74% |

We combined the original stable test set and the newly added dynamic test set into a new comprehensive test set comprising 49,200 sample points, and reorganized the confusion matrix according to the mapped predicted classes. In this matrix, the rows represent the map-predicted classes, and the columns represent the reference ground-truth classes:

| **Year** | $\mathbf{n_{RR}}$ | $\mathbf{n_{RN}}$ | $\mathbf{n_{R\cdot}}$ | $\mathbf{n_{NR}}$ | $\mathbf{n_{NN}}$ | $\mathbf{n_{N\cdot}}$ |
|---|---|---|---|---|---|---|
| 2015 | 21,935 | 2,448 | 24,383 | 2,665 | 22,152 | 24,817 |
| 2024 | 22,380 | 1,876 | 24,256 | 2,220 | 22,724 | 24,944 |

The area proportion matrix estimator is defined as:

$$\hat{p}_{hj} = W_h \cdot \frac{n_{hj}}{n_{h\cdot}}$$

Here, $\hat{p}_{hj}$ represents the estimated proportion of the area where the map predicted stratum $h$ and the reference ground-truth is class $j$, relative to the total study area. $\frac{n_{hj}}{n_{h\cdot}}$ denote the conditional probability within that stratum. The total area proportion of the reference ground-truth class $j$ is calculated by summing the area proportions of the true class $j$ across all strata. Based on this, the overall accuracy (OA) metric can be

derived:

$$\hat{OA}_w = \hat{p}_{RR} + \hat{p}_{NN}$$

Here, $\hat{OA}_w$ represents the area-weighted overall accuracy, indicating the proportion of the correctly classified area relative to the total area.

The formula for the standard error of the OA is:

$$S(\hat{OA}_w) = \sqrt{\sum_{h\in\{R,N\}} W_h^2 \cdot \frac{\hat{UA}_h(1-\hat{UA}_h)}{n_{h\cdot}-1}}$$

Here, $W_h$ represent the stratum area weight, $\hat{UA}_h$ is the user's accuracy for that stratum, and $n_{h\cdot}$ is the number of samples in that stratum.

Based on the above formulas and actual data, the accuracy metrics and their 95% confidence intervals (CI) are summarized in Table 6:

**Table 6. Area-weighted accuracy assessment results.**

| Year | Index | Balanced | Area-weighted | 95% CI |
|---|---|---|---|---|
| 2024 | OA | 91.67% | 91.17% | [90.84%, 91.50%] |
| 2015 | OA | 89.61% | 89.30% | [88.94%, 89.67%] |

The results show that the area-weighted overall accuracy (OA) is very close to the sample-level accuracy, and the 95% confidence interval is narrow. This indicates that at the overall pixel level, the classification quality of GlobalRice20 is robust and highly confident.

In summary, the area-weighted overall accuracy confirms the reliability of our classification quality; it remains stable before and after weighting and exhibits a narrow confidence interval, demonstrating that the regions mapped as rice have very high credibility. The pixel-level accuracy conclusions derived from the aforementioned supplementary experiments mutually complement the national-level area consistency

validation reported in the main text. Thank you once again for pointing out this methodological issue—which is highly critical yet easily overlooked in rare land-cover mapping at the global scale. This has allowed our accuracy assessment framework to better align with international good practice guidelines and has prompted us to provide a more transparent and rigorous explanation regarding the limitations of the reference sample design and the consistency of the area metrics.

**References**

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

**2. Representation of temporally missing optical observations**

**The manuscript states that missing data are encoded as "color." However, this appears to apply only when an entire data modality is unavailable. It is unclear how the proposed representation handles differences in the temporal density of optical observations.**

**In cloud-prone regions, optical observations may be unavailable for several consecutive months. When valid observations on either side of such a gap are connected by a line, the missing interval becomes visually indistinguishable from a continuous temporal transition. This issue may be particularly important in tropical rice-growing regions, where persistent cloud cover can create long gaps. A convolutional neural network could interpret the resulting straight segment as a genuine phenological increase or decrease rather than as an interval without observations.**

**The authors should clarify whether observation dates, temporal gaps, valid-observation masks, or the number of observations per compositing period are explicitly encoded in the input. They should also investigate whether these visually**

**concealed gaps can mislead the model and affect classification accuracy. An ablation experiment comparing the current representation with one that explicitly encodes temporal availability would help establish the robustness of the method.**

**RESPONSE:** Thank you very much for this valuable comment. We apologize that the original manuscript did not describe the treatment of temporally missing observations in the T2V representation with sufficient clarity, which may have led to the misunderstanding that "color" was used to encode missing data. We would first like to clarify that colors in the T2V images are not used to encode missingness; instead, fixed colors are assigned to distinguish the four remote-sensing variables, namely EVI, LSWI, VV, and VH. All samples are represented on a unified annual time axis (DOY 1–365), and each valid observation is mapped to its corresponding horizontal position according to its actual observation date and displayed using a marker. Therefore, although the final images do not explicitly display DOY values or coordinate ticks, the observation dates and differences in temporal sampling density among samples are still preserved through the spatial positions of valid observations along the fixed annual time axis. When the entire Sentinel-1 modality is unavailable, the corresponding SAR subplot is left blank.

We agree with the reviewer's concern regarding long gaps in optical observations. The current T2V representation does not additionally introduce a binary valid-observation mask, an observation-count channel, or a separate temporal-missingness indicator. When persistent cloud cover causes a long interval between two valid optical observations, the current representation still connects the valid observations on both sides of the gap according to their actual DOY positions. A long temporal interval is reflected by the greater horizontal distance between the two observations and by the absence of intermediate observation markers; however, the line connecting the two observations may still visually resemble a continuous temporal transition. Therefore, to directly investigate whether such visually concealed gaps could mislead the model and affect classification performance, we conducted an additional controlled ablation experiment on temporal-availability representation.

Specifically, using the 2024 dataset, we kept the EVI, LSWI, VV, and VH input variables, DOY coordinate range, image size, T2VRCM architecture, training and test samples, and training parameters identical, and changed only the way temporal gaps were rendered. The original Current T2V representation connects adjacent valid observations according to their actual DOY positions, even when missing intervals occur between them. As a comparison, we constructed a Gap-explicit T2V representation, in which valid observations are connected only within continuous temporal segments. When one or more temporal intervals are missing between two valid observations, no line is drawn across the gap and the corresponding interval is left blank. This design represents temporal availability more explicitly without introducing additional input channels or modifying the model architecture, thereby allowing us to directly evaluate the effect of connecting observations across missing intervals on classification performance.

**Table R6. Comparison between the current T2V representation and the gap-explicit temporal-availability representation.**

| Representation | UA (%) | PA (%) | F1 (%) | OA (%) |
|---|---|---|---|---|
| Current T2V | 92.52 | 92.12 | 92.32 | 92.33 |
| Gap-explicit T2V | 91.58 | 91.32 | 91.45 | 91.46 |

The experimental results show that, under the current dataset and experimental settings, explicitly disconnecting the curves across missing intervals did not improve classification performance. Compared with Gap-explicit T2V, Current T2V increased F1 from 91.45% to 92.32%, corresponding to an improvement of 0.87 percentage points, while OA increased from 91.46% to 92.33%. These results indicate that retaining connections between valid observations on a unified DOY time axis is more favorable for rice classification in the current task. This is mainly because directly connecting valid observations better preserves the overall annual curve shape and temporal

variation patterns associated with rice growth, whereas explicitly breaking the curves across missing intervals increases visual fragmentation and weakens the representation of long-term phenological information.

In addition, the T2V representation adopts separate optical and SAR subplots, with EVI and LSWI and Sentinel-1 VV and VH encoded on the same unified annual time axis, thereby preserving the temporal complementarity between the two types of remote-sensing information. Optical observations are susceptible to clouds and cloud shadows, whereas Sentinel-1 SAR provides all-weather and day-and-night observation capability. Therefore, when persistent cloud cover causes long gaps in the optical time series, VV and VH can still provide relatively continuous information on backscatter dynamics and offer additional constraints for rice phenology identification. Consequently, T2VRCM does not rely solely on the continuity of the optical curves for rice classification, but instead jointly exploits the complementary information provided by optical and SAR time series.

Overall, although the current T2V representation does not introduce an explicit valid-observation mask or observation-count channel, the unified DOY 1–365 time axis, together with the actual temporal positions and distribution of valid observations, still preserves information on observation dates and temporal sampling density. The additional ablation experiment further shows that the current representation, which maintains connections across missing intervals, achieves higher classification accuracy than the Gap-explicit T2V representation. This indicates that, for the task considered in this study, preserving the overall phenological curve structure is more beneficial for rice classification than explicitly disconnecting the curves across temporal gaps. Meanwhile, when optical observations become sparse because of persistent cloud cover, the Sentinel-1 VV and VH time series provide complementary backscatter information, thereby reducing the dependence of the classification process on the completeness of the optical time series.

Following your suggestion, we have further clarified the description of the T2V representation in the revised manuscript to help readers better understand how temporal

observations and missing intervals are represented. We sincerely thank the reviewer again for this valuable suggestion. The specific revision is provided below.

Page 10

Accordingly, the actual observation dates are explicitly retained through the DOY values of the valid observations, while missing temporal phases are represented by the absence of corresponding (time, value) tuples rather than by artificial observations. Thus, differences in temporal sampling density among samples are preserved in the original irregular time-series representation.

**3. Lack of a controlled comparison between representations**

**The current comparison changes both the input representation and the classification model. T2VRCM uses line-chart images with vision backbones, whereas the baseline approaches use raw time-series sequences with RF, LSTM, or Transformer models. Consequently, the observed performance difference cannot be attributed specifically to the image-based representation because model architecture and input representation vary simultaneously.**

**A more controlled experiment is needed to support the manuscript's explanation of why the proposed representation performs better. For example, the authors could use a common backbone, such as a Vision Transformer, and provide it with either image-based inputs or tokenized raw time-series inputs while keeping model capacity, training data, and optimization settings as comparable as possible. Alternatively, the same time-series model could be evaluated with different representations.**

**This concern does not dispute the empirical finding that T2VRCM is the best-performing model in the reported experiments. Rather, it asks the authors to provide a controlled comparison that isolates the contribution of the Time-Series-to-Vision representation.**

**RESPONSE:** Thank you very much for this valuable suggestion. We agree that, although the model comparison in the original Table 1 showed that T2VRCM achieved the highest classification accuracy under the current experimental settings, the native time-series models and the T2V image-based vision models differed simultaneously in both input representation and model architecture. Therefore, the performance improvement observed in the original cross-model comparison cannot be attributed solely to the Time-Series-to-Vision (T2V) representation. To address this issue, we conducted an additional controlled representation-level comparison using the 2024 dataset. It should be noted that the Stem, InceptionDWC, and FSFI modules in T2VRCM are specifically designed for two-dimensional image features; therefore, one-dimensional numerical time series cannot be directly fed into the identical T2VRCM architecture without modifying the input structure. To isolate the effect of the input representation as much as possible, we adopted the same Transformer encoder and classification head for comparison.

Specifically, in the first setting, EVI, LSWI, VV, and VH were directly represented as numerical time series, following the Transformer configuration used in Table 1. In the second setting, the same multi-source time series were converted into T2V images and fed into an identically configured Transformer encoder through image patch embedding. Both experiments were conducted using the 2024 dataset with the same training and test sets, while the Transformer depth, hidden dimension, number of attention heads, MLP configuration, classification head, and major training parameters were kept identical. Because numerical time-series representations and two-dimensional images have inherently different data structures, each setting used an embedding module appropriate for its input format, whereas the subsequent Transformer encoder and classification head remained unchanged. This design maximized the comparability of the downstream model architecture and training conditions while minimizing the influence of architectural differences on the comparison.

The experimental results are shown in Table R7. When using the numerical time-series representation, the Transformer achieved UA, PA, F1, and OA values of 85.87%, 85.03%, 85.45%, and 85.52%, respectively. After adopting the T2V image representation, the corresponding metrics increased to 88.05%, 87.77%, 87.91%, and 87.93%, respectively. In particular, F1 and OA improved by 2.46 and 2.41 percentage points, respectively. These results demonstrate that, under the same core Transformer encoder and training settings, the T2V image representation achieves higher classification accuracy and provides more direct experimental evidence for evaluating the contribution of the T2V representation itself.

**Table R7. Controlled comparison of numerical time-series and T2V image representations using the same Transformer encoder.**

| Representation | Model | UA (%) | PA (%) | F1 (%) | OA (%) |
|---|---|---|---|---|---|
| Numerical time series | Transformer | 85.87 | 85.03 | 85.45 | 85.52 |
| T2V image | Transformer | 88.05 | 87.77 | 87.91 | 87.93 |

The observed performance improvement indicates that the structured organization of irregular multi-source time series provided by T2V is beneficial for enhancing the model's feature representation capability. In the numerical time-series representation, information is primarily provided to the model as values distributed across different time steps and variables. In contrast, T2V organizes the actual temporal positions, overall curve shapes, variations across different growth stages, and the relationships among EVI, LSWI, VV, and VH within a unified two-dimensional space. This enables the model to jointly exploit temporal positions, multivariable variation patterns, and overall phenological curve structures for feature extraction. Therefore, T2V not only changes the form of the input representation, but also reorganizes temporal and multivariable information that is originally distributed across different time steps and variables into two-dimensional visual patterns with explicit structural relationships,

thereby facilitating the extraction of discriminative features associated with rice phenological processes.

This controlled experiment also complements the original ablation study. The newly added experiment primarily evaluates the contribution of the input representation to classification performance, whereas the ablation experiments on InceptionDWC and FSFI in the original Table 2 further evaluate the contribution of the T2VRCM network architecture. Together, these experiments provide a clearer distinction between the effects of the T2V visual representation and the T2VRCM network architecture on the final classification performance. Following your suggestion, we have added this controlled comparison and the corresponding analysis to Discussion Section 5.3 of the revised manuscript and further clarified the respective contributions of the T2V representation and the T2VRCM network architecture to classification performance. We sincerely thank the reviewer again for this valuable suggestion.

The specific revisions are provided below.

**Pages 30-31**

5.3 Effects of the Time-Series-to-Vision Representation on Classification Performance

To further distinguish the contribution of the T2V visual representation from that of the subsequent network architecture, we conducted a controlled representation-level comparison based on the 2024 dataset. Because the Stem, InceptionDWC, and FSFI modules in T2VRCM are specifically designed for two-dimensional image features, one-dimensional numerical time series cannot be directly fed into the identical T2VRCM architecture without modifying the input structure. Therefore, the same Transformer encoder and classification head were adopted for comparison. The first setting used numerical time-series representations of EVI, LSWI, VV, and VH, following the preprocessing procedure and Transformer baseline configuration used in Table 1. The second setting converted the same multi-source time series into T2V images and fed them into the same Transformer encoder through image patch

embedding. The two experiments used exactly the same training and test sets and maintained identical Transformer depth, hidden dimension, number of attention heads, classification head, and training parameters, thereby minimizing the influence of differences in network architecture and training settings on the comparison.

The experimental results are shown in Table 5. Using the numerical time-series representation, the Transformer achieved UA, PA, F1, and OA values of 85.87%, 85.03%, 85.45%, and 85.52%, respectively. After adopting the T2V image representation, the corresponding metrics increased to 88.05%, 87.77%, 87.91%, and 87.93%. Under the same core Transformer encoder and major training settings, the T2V representation improved F1 and OA by 2.46 and 2.41 percentage points, respectively, demonstrating that transforming multi-source time series into a two-dimensional visual representation can further improve rice classification performance.

**Table 5. Comparison of numerical time-series and T2V image representations using the same Transformer encoder.**

| Representation | Model | UA (%) | PA (%) | F1 (%) | OA (%) |
|---|---|---|---|---|---|
| Numerical time series | Transformer | 85.87 | 85.03 | 85.45 | 85.52 |
| T2V image | Transformer | 88.05 | 87.77 | 87.91 | 87.93 |

The performance improvement mainly stems from the way T2V reorganizes irregular multi-source time-series structures. Specifically, T2V maps the actual temporal positions of different remote-sensing variables, the overall annual curve shapes, stage-wise variations, and the relationships among EVI, LSWI, VV, and VH into a unified two-dimensional visual space, thereby transforming information originally distributed across different time steps and variables into an integrated representation with explicit structural relationships. Based on this representation, the model can simultaneously exploit temporal positions, multivariable variation relationships, and overall phenological curve characteristics for classification, enabling more effective extraction of structural information associated with rice growth processes. Although converting

continuous numerical time series into images inevitably introduces some loss of numerical precision, the experimental results show that this loss does not offset the advantages provided by the structured visual representation, ultimately resulting in an overall improvement in classification accuracy.

**4. Cross-sensor comparability between 2015 and 2024**

**The 2015 map is based on Landsat 8, whereas the 2024 map uses Sentinel-2. Nevertheless, LSWI and EVI are calculated using the same nominal formulas for both years. Landsat 8 and Sentinel-2 have different band definitions, bandwidths, central wavelengths, and spectral response functions. As a result, indices with the same name may have systematic sensor-dependent differences.**

**The manuscript does not appear to describe any cross-sensor calibration or harmonization. Without such treatment, the 2015 and 2024 products may not be directly comparable, which has important implications for both the change analysis and the reported accuracy of the 2015 map.**

**The authors should explain how cross-sensor consistency was established. Relevant procedures might include spectral band adjustment, radiometric normalization, sensor-specific index transformations, harmonization using overlapping observations, or distributional comparisons over stable reference regions. At minimum, the manuscript should present an empirical assessment showing whether Landsat 8- and Sentinel-2-derived EVI and LSWI values are comparable in locations and periods for which observations from both sensors are available.**

**RESPONSE:** Thank you very much for this valuable suggestion. We agree that Landsat-8 and Sentinel-2 differ in terms of band configuration, central wavelength, bandwidth, and spectral response functions. Therefore, even when EVI and LSWI are calculated using the same nominal formulas, systematic sensor-dependent differences may still exist between the index values derived from the two sensors. We apologize

that the original manuscript did not sufficiently describe the model-training strategy for 2015 and 2024, which may have led readers to assume that a single model was directly applied across the two sensor configurations. Following your comment, we have further clarified this issue and conducted an additional analysis.

First, we would like to clarify that the GlobalRice20 products for 2015 and 2024 were generated using separately trained year-specific T2VRCM models. The 2015 model was trained using Landsat-8 optical data, whereas the 2024 model was trained using Sentinel-2 optical data. A single model was not directly applied across the two sensors. In addition, no additional cross-sensor calibration or harmonization between Landsat-8 and Sentinel-2 was performed in this study, including spectral band adjustment, radiometric normalization, or sensor-specific index transformation. Therefore, we do not assume that EVI and LSWI derived from the two sensors have identical numerical distributions, nor were indices from the two sensors directly mixed within a single model. The purpose of year-specific training was to allow each model to adapt to the feature distributions associated with the corresponding sensor and annual observation conditions, thereby avoiding the effects associated with directly applying an unadapted model across different sensor domains.

To further examine classification performance under the two optical data sources and their cross-domain transferability, we conducted an additional Optical-only cross-domain transfer experiment. Specifically, we first evaluated models trained and tested independently using the 2015 Landsat-8 data and the 2024 Sentinel-2 data, respectively. We then directly applied the Optical-only model trained on the 2024 Sentinel-2 data to the 2015 Landsat-8 data without retraining or sensor adaptation.

**Table R8. Classification performance under within-domain and cross-domain optical-data settings.**

| Training data | Testing data | UA (%) | PA (%) | F1 (%) | OA (%) |
|---|---|---|---|---|---|
| 2015 Landsat-8 | 2015 Landsat-8 | 84.93 | 86.00 | 85.46 | 85.37 |
| 2024 Sentinel-2 | 2024 Sentinel-2 | 86.76 | 87.62 | 87.19 | 87.13 |
| 2024 Sentinel-2 | 2015 Landsat-8 | 80.07 | 80.55 | 80.31 | 80.25 |

The results show that, when the models were trained and tested within their corresponding year-specific sensor domains, the Optical-only F1 values were 85.46% for the 2015 Landsat-8 data and 87.19% for the 2024 Sentinel-2 data. When the model trained on the 2024 Sentinel-2 data was directly applied to the 2015 Landsat-8 data, F1 decreased to 80.31%, representing a reduction of 6.88 percentage points relative to the within-domain 2024 Sentinel-2 evaluation and 5.15 percentage points relative to the model specifically trained on the 2015 Landsat-8 data. These results indicate that directly transferring a model across year–sensor domains without adaptation leads to a clear decline in classification performance, reflecting non-negligible distributional differences between the two data domains. This finding further supports the use of separately trained year-specific models. For the production and accuracy assessment of each annual product, the model was independently trained and tested using data from the corresponding year and sensor, thereby avoiding the direct use of Landsat-8 and Sentinel-2 features within the same unadapted model.

Following your suggestion, we have further clarified in the Methods section that separate models were trained and applied for 2015 and 2024. We sincerely thank the reviewer again for this valuable suggestion. The specific revision is provided below.

**Page 16**

Considering the differences in spectral characteristics and temporal sampling between Landsat-8 and Sentinel-2, separate year-specific T2VRCM models were trained for 2015 and 2024 rather than directly applying a single model across the two sensor configurations.

**5. Language and presentation**

**The manuscript contains numerous expressions that are non-idiomatic, overly promotional, or insufficiently precise for a scientific paper. Examples include phrases such as "tracking SDG 2 progress" and "the transformation strategy cleverly encodes." The latter is particularly subjective and should be replaced with neutral technical language.**

**Several sentences contain multiple semicolons or overly complex grammatical structures, and the formatting of semicolons is inconsistent. Some terminology should also be reconsidered. For example, although "missingness" is used in some technical literature, clearer expressions such as "data absence," "missing observations," or "patterns of missing data" may be more appropriate depending on the intended meaning.**

**The manuscript would benefit from another thorough round of professional English-language editing. The revision should focus not only on grammar but also on scientific precision, consistency of terminology, sentence structure, punctuation, and the removal of promotional or subjective wording.**

**RESPONSE:** We thank the reviewer for this valuable comment. In response, we have carefully revised the language throughout the manuscript to further improve its clarity, accuracy, and consistency in scientific expression. The revisions mainly include removing or softening subjective or overly promotional wording, standardizing and harmonizing technical terminology, simplifying overly complex sentence structures, ensuring consistent punctuation usage, and conducting a comprehensive check and

revision of grammar and wording throughout the manuscript. The specific revision is provided below.

**Page 1**

Accurate, high-resolution spatial data of paddy rice are indispensable for assessing global food security and supporting assessments related to Sustainable Development Goal 2 (Zero Hunger).

**Page 1**

Cross-comparison with national agricultural statistics reveals a high coefficient of determination ($R^2$ =0.91 for 2024), indicating good agreement at the national scale.

**Page 3**

In addition, fusing optical and SAR modalities is complicated by temporal asynchrony. The different revisit cycles of Landsat, Sentinel-2, and Sentinel-1 result in irregularly sampled time series, which pose challenges for traditional models.

To bridge this gap, we developed the GlobalRice20 dataset, the first globally consistent 20 m resolution paddy rice map for the years 2015 and 2024. To address irregular sampling and missing observations or modalities, we developed a Time-series-to-Vision framework that transforms heterogeneous optical and SAR time series into standardized 2D visual representations. Unlike conventional methods that rely on regularly sampled time series, this framework enables the tailored vision model (T2VRCM) to effectively capture rice phenological patterns under incomplete and irregular observations. In this study, we generated, validated, and analyzed the GlobalRice20 dataset, a 20-m resolution global rice map covering 98 major rice-producing countries for 2015 and 2024. Using a Time-series-to-Vision harmonization strategy to address cloud contamination, asynchronous sensor observations, and missing modalities, the resulting maps achieved an overall accuracy of 92.33% and showed strong agreement with national statistics, indicating the reliability and applicability of the dataset at the global scale. Furthermore, we demonstrate the

dataset's utility for monitoring progress toward SDG 2 by quantifying key decadal changes, such as the 15.7% expansion of rice cultivation in Africa, highlighting its potential to support analyses of agricultural change and food security.

**Page 7**

These include APRA500 (Han et al., 2022), a 500-m product mapping cropping intensity across 21 Asian countries using MODIS data, and SingleSeasonRice-China (Shen et al., 2023b), a 10/20-m Sentinel-based dataset covering 21 Chinese provinces. CCD-Rice (Shen et al., 2025) is a long-term (1990–2016) 30-m dataset for China's eastern monsoon region, while ChinaCropArea1km (Luo et al., 2020) and SPAM-China (Dai et al., 2025) are two 1-km products generated from GLASS LAI and disaggregated statistics, respectively, and were used as coarse-resolution baselines. Open-SEA-Rice-10 (Ginting et al., 2025) is a 10-m dataset that distinguishes triple-cropping systems in Southeast Asia, whereas AsiaRiceYield4km (Wu et al., 2022) is a long-term (1995–2015) 4-km gridded product developed using machine learning to estimate seasonal rice yields across Asia. GCD-Rice (Li et al., 2025) provides a long-term (1990–2023) 30-m dataset of seasonal rice distribution across 16 Asian countries based on Landsat and Sentinel-1 data. The Japan Historical Rice Map (Carrasco et al., 2022) maps rice field distribution in Japan from 1985 to 2019 using Landsat imagery, phenological information, and temporal aggregation, while the South Korean Dynamic Rice Map (Jo et al., 2023) uses a Recurrent U-Net model for dynamic paddy rice mapping. NESEA-Rice (Han et al., 2021) provides annual 10-m paddy rice maps for Northeast and Southeast Asia from 2017 to 2019 by integrating Sentinel-1 and MODIS data. The USDA Cropland Data Layer (Boryan et al., 2011) is a 30-m crop land-cover product for the United States generated by the USDA using decision-tree classification of satellite imagery. Together, these datasets span spatial resolutions from 10 m to 4 km and include both single- and multi-cropping systems, providing reference products for assessing the spatial consistency of GlobalRice20 across different scales.

**Page 11**

The Time-Series-to-Vision transformation converts heterogeneous time-series data with irregular sampling and missing observations into standardized two-dimensional image representations.

**Page 13-14**

This design enables the block to extract discriminative phenological features from the 2D time-series graphs while reducing sensitivity to missing observations and noise.

**Page 15**

To evaluate the performance of T2VRCM and the time-series-to-vision strategy, we compared the proposed model with two categories of baseline models.

**Page 16**

T2VRCM achieved an OA of 92.33% and an F1 score of 92.32%, surpassing the best-performing baseline (MambaOut-Femto) by margins of 2.89% and 2.96%, respectively.

**Page 27**

Africa showed the largest relative increase among the continents, with mapped rice area increasing by 15.7%.

## Other suggested modifications

**1. The optical temporal sampling is inconsistent between sections. Section 2.2 describes only cloud masking and index calculation, with no compositing, while Section 3.1 mentions 5-day composites for Sentinel-2. Please state the actual optical sampling scheme so the method can be reproduced.**

**RESPONSE:** We sincerely thank the reviewer for pointing out this inconsistency. We agree that the previous description in Section 2.2 did not clearly state the optical temporal compositing scheme, which could reduce reproducibility and cause confusion

with Section 3.1. In the revised manuscript, we have clarified that the optical data were temporally composited according to the nominal revisit cycle of each sensor. Specifically, Sentinel-2 observations in 2024 were composited at a 5-day interval, while Landsat-8 observations in 2015 were composited at a 16-day interval. LSWI and EVI were then calculated for each optical composite after cloud masking. We have also revised Section 3.1 to ensure consistency across the manuscript. Once again, we sincerely appreciate your valuable suggestion.

The specific revisions are as follows:

**Page 5**

For 2024, we utilized Sentinel-2 MSI Level-2A surface reflectance data (10-m resolution), applying a cloud-masking algorithm to remove invalid observations. The valid Sentinel-2 observations were then composited at a 5-day interval. For the 2015 target year, as Sentinel-2 data were not yet available, we utilized Landsat-8 OLI surface reflectance data with a 30-m spatial resolution. After cloud masking, the valid Landsat-8 observations were composited at a 16-day interval.

**Page 10**

(1) Inconsistent revisit cycles: Different sensors have varying revisit cycles (e.g., 5-day composites for Sentinel-2, 12-day composites for Sentinel-1 and 16-day composites for Landsat-8);

**2. The non-rice class is undocumented. The 1:1 balance is stated, but not what land covers the negative samples represent (other crops, water, forest, urban...) or how they were distributed. For a data paper, a short table or figure of non-rice class proportions would help readers a lot.**

**RESPONSE:** Thank you for your valuable feedback. We fully agree that for a data descriptor paper, merely stating a 1:1 ratio for the rice/non-rice samples without detailing the specific land cover composition of the negative (non-rice) samples indeed

hinders readers from comprehensively evaluating the representativeness and reliability of the reference dataset.

In response to this comment, we have supplemented the detailed land cover composition of the non-rice samples. We conducted a spatial overlay analysis of all approximately 82,000 non-rice sample points with the 2024 Esri 10 m Annual Land Cover product (Karra et al., 2021). Through this analysis, we extracted the land cover class corresponding to each sample point (e.g., Water, Trees, Flooded Vegetation, Crops, Built Area, Bare Ground, and Rangeland) and calculated the count and proportion of each category within the non-rice samples. Because this study adopted a "stable sample" strategy when constructing the reference samples, meaning that all sample points maintained a consistent land cover type between 2015 and 2024, the 2024 land cover annotations can reasonably represent the land cover attributes of these sample points in 2015. Based on this premise, a single overlay analysis is sufficient to characterize the sample composition for both target years. The non-rice sample points are distributed across the same 98 countries/regions and climate zones as the rice samples, ensuring that the negative samples have comparable spatial and geographical coverage to the positive samples rather than being concentrated in isolated regions.

**Table R9. Land-cover composition of the non-rice reference samples.**

| Land-Cover Class | Sample Count | Proportion (%) |
|---|---|---|
| Crops (non-rice) | 37471 | 45.70 |
| Water | 15374 | 18.75 |
| Flooded Vegetation | 6283 | 7.66 |
| Trees / Forest | 11386 | 13.89 |
| Built Area / Urban | 5961 | 7.27 |
| Others | 5525 | 6.74 |
| Total | 82,000 | 100.00 |

The results indicate that the non-rice samples encompass other crops (45.70%), water (18.75%), flooded vegetation (7.66%), trees/forests (13.89%), built areas (7.27%), and other minor categories (6.74%) (see Table R9). Notably, other crops account for a substantial 45.70%, while water and flooded vegetation together account for approximately 26.41%. These land cover types are particularly relevant for distinguishing paddy rice because other crops may exhibit similar phenological patterns, while water and flooded vegetation can show optical and SAR characteristics similar to those of flooded rice fields during certain growth stages. Their substantial representation within the negative samples therefore helps ensure that the reference dataset contains challenging non-rice cases and supports the model in learning more discriminative boundaries between rice and non-rice land covers.

**References**

Karra, K., Kontgis, C., Statman-Weil, Z., Mazzariello, J. C., Mathis, M., and Brumby, S. P.: Global land use/land cover with Sentinel 2 and deep learning, 2021 IEEE international geoscience and remote sensing symposium IGARSS, 4704-4707, 2021.

**3. Samples were built from visual interpretation process integrating high-resolution optical imagery and existing regional rice maps, but the manuscript does not say which came from products and which from interpretation. Please separate the two so readers know how much of the reference is truly independent.**

**RESPONSE:** Thank you for your valuable feedback. We fully agree with your perspective; if the labels for a portion of the reference samples were directly derived from existing data products, the independence of these samples when subsequently used for accuracy assessment would indeed be questionable. We apologize if our explanation of the sample sources in the manuscript was not sufficiently clear.

To clarify, the entire set of 164,000 reference samples in this study does not consist of a mixture of product-derived and interpretation-derived samples. Instead, all samples

were obtained exclusively through manual visual interpretation. None of the sample labels were directly extracted from existing regional paddy rice products.

Specifically, after selecting spatially discrete candidate locations within the rice-cultivating regions of each country, the annotators utilized the annual high-resolution optical imagery corresponding to the target years (2015 and 2024) as the primary basis for determining the final land cover label for each sample. Existing regional rice maps served purely as auxiliary references during the annotation process. For instance, they assisted annotators in roughly delineating potential rice cultivation extents and provided background information or cross-verification for challenging plots. However, they did not directly determine or substitute the final manual interpretation results. In other words, the existing products functioned solely as supplementary information rather than as the source of the labels in our annotation workflow.

This clarification also bears profound significance: because the label generation process for the reference samples was entirely independent of existing regional rice products, there is no risk of circular validation between the reference samples and the comparison benchmarks when we utilize these products and national statistical data for spatial cross-comparison and statistical consistency verification of GlobalRice20 (as detailed in Sections 2.3.2 and 4.3.3). This strictly guarantees the independence and credibility of our overall validation framework.

We hope this explanation resolves any misunderstandings. Thank you once again for raising this important point, which has allowed us to further clarify and strengthen the descriptions regarding label independence in our methodology section.

The specific revisions are as follows.

**Page 6-7**

All samples were derived through rigorous visual interpretation of discrete locations distributed across rice-cultivating regions in each country, rather than being sampled from continuous spatial blocks. Consequently, this design mitigates the likelihood of adjacent or duplicate samples appearing in both the training and test sets, thereby

reducing the risk of accuracy overestimation driven by spatial autocorrelation.

The sample annotation primarily relied on high-resolution optical imagery corresponding to the target years (2015 and 2024), including Google Earth, Landsat-8, and Sentinel-2 imagery. Additionally, existing regional rice maps were incorporated to locate potential cultivation extents. The annotation process was collaboratively executed by a dedicated team comprising 8 graduate students, 4 professors with extensive experience in agricultural remote sensing, and 1 expert from the Academy of Agricultural Sciences. This process involved multiple stages: initially, the expert provided unified guidance and established standardized visual interpretation rules for the team; subsequently, the 8 graduate students conducted preliminary labeling of candidate samples based on these criteria; finally, the 4 professors comprehensively audited the preliminary results. Any discrepancies encountered during the review were resolved through multiple team consensus meetings, thereby ensuring the high accuracy and reliability of the final labels.

**4. Some products used to build the samples reappear in the Section 4.3.2 comparison. Please confirm whether any product was used heavily in both steps, to rule out circular validation.**

**RESPONSE:** Thank you for your valuable feedback. We fully agree with your perspective that if the products used for spatial cross-comparison were heavily utilized during the sample construction phase, the comparison results would lose their independence, resulting in circular validation. Here, we wish to address this issue candidly and explicitly: we do not deny that the existing regional rice maps used as auxiliary references in Section 2.3.1 overlap with some of the comparison products listed in Section 4.3.2. This is not surprising, as both target highly overlapping study areas and data sources, and the availability of existing public high-resolution regional rice products is inherently limited. However, it must be emphasized that the repeated use of these products does not introduce the risk of circular validation that you are

concerned about. The specific reasons are as follows:

First, the label generation process is entirely independent of existing products. All reference samples were determined exclusively through manual visual interpretation, with the core interpretation basis being high-resolution optical imagery corresponding to the target year (2015 or 2024). Existing regional rice maps provided only background reference information during this process, such as assisting annotators in roughly locating potential rice-cultivating areas. Annotators were explicitly not allowed to directly extract the classification results of any product as sample labels. All annotation results were individually reviewed and, when necessary, corrected by professors with experience in rice mapping. Therefore, the labels were controlled by manual interpretation, not by any existing products.

Second, the involvement of existing products during the annotation phase was extremely limited. Even though certain existing products served as references during annotation, their role was confined to coarse-grained paddy localization—a non-decisive auxiliary function rather than a source of labels. We did not establish any protocol where product results equated to labels for any specific product; the core of the annotation work remained point-by-point manual interpretation based on remote sensing imagery.

Third, the functional roles played by existing products in these two stages are fundamentally different. During the sample construction phase (Section 2.3.1), the role of existing products was to provide background reference for annotators, serving an auxiliary interpretation function. Conversely, in Section 4.3.2, these products acted as completely independent external benchmarks to evaluate the consistency and relative advantages between GlobalRice20 and other products from macro perspectives such as spatial distribution patterns and boundary details. The two applications differ entirely in their functional objectives and how the information is utilized, precluding any circular validation.

Finally, the overall validation framework of this study is safeguarded by multiple

layers and sources. In addition to the spatial cross-comparison in Section 4.3.2, this study also conducted an accuracy assessment based on independent test samples in Section 4.3.1, and a national-level area consistency validation based on official USDA and FAO statistics in Section 4.3.3. These cross-validate the reliability of GlobalRice20 from three mutually independent dimensions: the sample level, the product level, and the statistical level. This multi-layered validation architecture further mitigates the impact of potential biases in any single stage on the overall conclusions and eliminates the risk of circular validation by design.

We have added relevant text to Section 2.3.1 of the revised manuscript to explicitly clarify the auxiliary role of existing products during the annotation phase.

The specific revisions are as follows.

**Page 6-7**

The sample annotation primarily relied on high-resolution optical imagery corresponding to the target years (2015 and 2024), including Google Earth, Landsat-8, and Sentinel-2 imagery. Additionally, existing regional rice maps were incorporated to locate potential cultivation extents. The annotation process was collaboratively executed by a dedicated team comprising 8 graduate students, 4 professors with extensive experience in agricultural remote sensing, and 1 expert from the Academy of Agricultural Sciences. This process involved multiple stages: initially, the expert provided unified guidance and established standardized visual interpretation rules for the team; subsequently, the 8 graduate students conducted preliminary labeling of candidate samples based on these criteria; finally, the 4 professors comprehensively audited the preliminary results. Any discrepancies encountered during the review were resolved through multiple team consensus meetings, thereby ensuring the high accuracy and reliability of the final labels.

**5. The line-chart rendering is under-specified: no color values, line width, Y-axis rule (fixed vs. per-sample), resize method, or axis drawing. Since a CNN reads**

**pixel patterns, these details matter for reproducibility. Please add a short implementation note or release the plotting script.**

**RESPONSE:** Thank you very much for this valuable suggestion. We agree that the specific rendering parameters of the line charts directly affect the final pixel patterns of the T2V images, and that these implementation details are therefore important for the reproducibility of the proposed method. We apologize that the original manuscript did not provide sufficiently detailed descriptions of the rendering parameters, including color settings, line width, coordinate ranges, marker styles, and image size. Following your suggestion, we carefully re-examined the plotting code that was actually used to generate the training and testing input images and have added the relevant implementation details to the revised manuscript.

Specifically, each sample was directly rendered as a 224 × 224 pixel uint8 RGB image, without any scaling or resampling from an intermediate image size. The image consists of two vertically stacked panels of equal height, with each panel occupying 112 pixels. The upper panel displays EVI and LSWI, while the lower panel displays Sentinel-1 VV and VH. In both panels, the horizontal axis is fixed to the DOY range 1–365, and all valid observations are linearly mapped to their corresponding horizontal pixel positions according to their actual DOY values. The Y-axis is also fixed rather than automatically scaled for each sample. Specifically, the value range of the optical panel is fixed to −1.0 to 1.0, while that of the SAR panel is fixed to −40.0 to 5.0 dB. Therefore, for all samples, the same feature value corresponds to the same vertical pixel position, ensuring a consistent mapping between numerical values and pixel locations across samples.

The four remote-sensing variables are rendered using fixed colors: EVI is shown in #008000 (RGB 0, 128, 0), LSWI in #0081CF (RGB 0, 129, 207), VV in #4E8397 (RGB 78, 131, 151), and VH in #C34A36 (RGB 195, 74, 54). All curves are rendered as 1-pixel anti-aliased solid lines. Each valid observation is marked with a five-pointed star of 2-pixel radius, filled in yellow (#FFFF00), with the outline color matching the corresponding curve and an outline width of 1 pixel. Invalid observations identified by

upstream cloud and data-quality masks are uniformly encoded as −9999 during export and are filtered out together with non-finite values before plotting; the numerical value 0 is not treated as a missing-value indicator, and valid zero-valued observations are retained. The retained valid observations are sorted in ascending order of DOY before plotting and then connected sequentially.

The image background is fixed to white, and both panels are enclosed by 1-pixel black frames. No axis ticks, DOY numbers, or textual labels are displayed in the final image. Only the fixed spatial coordinate mapping, time-series curves, valid-observation markers, and panel borders are retained. Because the T2V images are rendered directly at the final resolution of 224 × 224 pixels, no additional resizing or resampling is involved in the entire generation process. Following your suggestion, we have added the above rendering parameters and implementation details to the revised manuscript to improve the reproducibility of the T2V representation. We sincerely thank the reviewer again for this valuable suggestion.

The specific revisions are provided below.

**Page 10-11**

The Y-axis ranges are fixed to −1.0–1.0 for the optical subplot and −40.0–5.0 dB for the SAR subplot, with no sample-specific auto-scaling.

To enable the model to distinguish between the different lines, four distinct yet consistent colors are assigned to the four features throughout the dataset, with EVI, LSWI, VV, and VH represented by #008000, #0081CF, #4E8397, and #C34A36, respectively.

Finally, the two subplots are stacked vertically and directly rendered as a standard RGB image $X_i$ with dimensions 224 × 224 pixels. All curves are rendered as 1-pixel anti-aliased solid lines, and valid observations are marked using yellow-filled five-pointed stars with a radius of 2 pixels and outlines matching the corresponding curve colors. Axis ticks and textual labels are not displayed, while a white background and 1-pixel

black frames are retained for both subplots.

**6. The train/test split is not described, nor whether models were trained per year or combined. With strong spatial autocorrelation, random splits can overstate accuracy. Please state the protocol, and ideally add a leave-region-out check.**

**RESPONSE:** Thank you for your valuable questions regarding the training/test partition protocol, the annual modeling approach, and the issue of spatial autocorrelation. We address each of them below:

(1) Regarding the training/test partition protocol and independent annual modeling

The relevant content has been fully supplemented in Section 2.3.1 (Reference Sample Set) of the revised manuscript. The 164,000 reference samples were spatially stratified with countries acting as the spatial stratification units. Within each country, spatially stratified random partitioning was conducted, maintaining a fixed 4:1 ratio for the training and test sets (yielding approximately 131,200 training samples and 32,800 test samples globally). A 1:1 class balance between rice and non-rice samples was preserved in both subsets. For the two target years of 2015 and 2024, we independently trained two separate T2VRCM models. Each model was trained and evaluated using only the remote sensing time-series features corresponding to its respective year. There is no instance of combining samples from both years to train a single model, nor is there any cross-year data leakage.

(2) Regarding the risk of accuracy overestimation caused by spatial autocorrelation

We fully understand your concern regarding spatial autocorrelation, which is a common issue requiring careful consideration in remote sensing mapping. Given the specific sample design of this study, we believe this risk has been effectively mitigated for the following reasons:

First, as detailed in Section 2.3.1, all reference samples in this study were obtained through manual visual interpretation of spatially discrete locations distributed across

rice-cultivating regions in each country, rather than being sampled from spatially contiguous blocks. This point-based, discrete sampling approach inherently ensures substantial spatial intervals between sample points from the very source of sample generation, significantly reducing the likelihood of adjacent or duplicate pixels appearing simultaneously in both the training and test sets.

Second, we employed spatially stratified random partitioning using countries as units, rather than a globally unconstrained random shuffle. This design not only guarantees that each country has representative samples in both the training and test sets but also ensures that the training and test samples within the same country originate from discrete, independent geographic locations rather than adjacent pixels within the same contiguous patch. This further controls the risk of spatial autocorrelation leakage at a methodological level.

Our sample partitioning strategy can be summarized as a combination of spatially discrete site selection and country-level stratified random partitioning. This is a prevalent practice in current global- and regional-scale agricultural remote sensing mapping, effectively controlling spatial autocorrelation risks and remaining consistent with established conventions.

(3) Regarding the suggestion to add a leave-region-out validation

To rigorously address your concerns regarding spatial autocorrelation and to test the spatial transferability of our method, we conducted a strict leave-region-out validation. Specifically, utilizing the 2024 dataset, we trained the model using all reference samples while explicitly excluding samples from China. Subsequently, we tested the model's performance strictly on these unseen samples within China. We have added a new table (Table R10) to compare the leave-region-out performance of different models. The baseline models are identical to those evaluated in the original manuscript.

**Table R10. Leave-Region-Out Performance Comparison of Different Models.**

| Model | UA | PA | F1 | OA |
|---|---|---|---|---|
| RF | 81.37% | 80.24% | 80.81% | 80.93% |
| LSTM | 84.09% | 82.51% | 83.29% | 83.45% |
| Transformer | 83.74% | 84.13% | 83.94% | 83.90% |
| ResNet-18 | 85.91% | 84.86% | 85.39% | 85.47% |
| ViT-Tiny | 86.33% | 86.67% | 86.50% | 86.47% |
| Swin-Tiny | 87.16% | 86.83% | 87.00% | 87.02% |
| ConvNeXt-Atto | 87.42% | 86.71% | 87.07% | 87.12% |
| MambaOut-Femto | 88.63% | 87.49% | 88.06% | 88.13% |
| T2VRCM(Ours) | 91.24% | 90.58% | 90.91% | 90.94% |

As shown in Table R10, under the strict leave-region-out setting, the accuracy of the models experienced a slight decline compared to the random partitioning results, which corroborates your point regarding the impact of spatial autocorrelation. However, our proposed T2VRCM demonstrated robust spatial transferability, maintaining a high overall accuracy (OA) of 90.94% and significantly outperforming other vision models and native time-series models. This indicates that our "Time-Series-to-Vision" framework, coupled with the dynamic gating and Full-Stage Feature Interaction (FSFI) mechanisms, successfully extracts the global phenological features of paddy rice, rather than merely overfitting to local spatial patterns.

In summary, we believe the existing sample design has effectively controlled the risk of accuracy overestimation caused by spatial autocorrelation. Furthermore, the existing national-level statistical validation and multi-product spatial cross-comparisons have thoroughly verified the model's geographic generalization capability from external, independent perspectives. Combined with the supplementary leave-region-out independent validation results, we have further proven—from both the algorithmic foundation and empirical perspectives—that the T2VRCM model can accurately extract universal rice phenological features, rather than merely fitting or relying on the spatial autocorrelation of localized regions. Therefore, the GlobalRice20 dataset generated in this study exhibits high reliability and spatial consistency across diverse

geographical environments and agricultural systems, fully meeting the requirements for large-scale agricultural monitoring and Sustainable Development Goals (SDGs) assessment.

The specific revisions are as follows.

**Pages 6-7**

To train and validate the T2VRCM model, a comprehensive reference dataset was constructed containing 164,000 samples across 98 countries. All samples were derived through rigorous visual interpretation of discrete locations distributed across rice-cultivating regions in each country, rather than being sampled from continuous spatial blocks. Consequently, this design mitigates the likelihood of adjacent or duplicate samples appearing in both the training and test sets, thereby reducing the risk of accuracy overestimation driven by spatial autocorrelation.

The sample annotation primarily relied on high-resolution optical imagery corresponding to the target years (2015 and 2024), including Google Earth, Landsat-8, and Sentinel-2 imagery. Additionally, existing regional rice maps were incorporated to locate potential cultivation extents. The annotation process was collaboratively executed by a dedicated team comprising 8 graduate students, 4 professors with extensive experience in agricultural remote sensing, and 1 expert from the Academy of Agricultural Sciences. This process involved multiple stages: initially, the expert provided unified guidance and established standardized visual interpretation rules for the team; subsequently, the 8 graduate students conducted preliminary labeling of candidate samples based on these criteria; finally, the 4 professors comprehensively audited the preliminary results. Any discrepancies encountered during the review were resolved through multiple team consensus meetings, thereby ensuring the high accuracy and reliability of the final labels.

The sample points were partitioned into training and test sets employing a spatially stratified random sampling strategy, utilizing each country as a stratification unit. Within each country, samples were allocated to the training and test sets at a fixed ratio

of 4:1 (yielding approximately 131,200 training samples and 32,800 test samples globally). A balanced 1:1 ratio between rice and non-rice samples was maintained in both subsets to prevent class imbalance.

Crucially, to ensure temporal consistency and reduce label noise, we adopted a "stable sample" strategy: selected samples were restricted to locations that maintained a consistent land cover type (either persistent rice or persistent non-rice) in both 2015 and 2024. Consequently, identical sample locations and training/test splits were applied for both 2015 and 2024. However, we independently trained and evaluated two separate T2VRCM models for the two target years, utilizing the time-series features corresponding to each respective year. The training and evaluation of each model were strictly confined to the contemporary training and test sets of that specific year.

**7. The cross-comparison with existing maps is entirely qualitative. Even the US check against USDA CDL reports only "high agreement" with no numbers. Please add at least one quantitative metric against a solid reference like CDL.**

**RESPONSE:** Thank you for this valuable suggestion. We fully agree that the current cross-comparison is mainly qualitative and does not provide sufficient quantitative evidence. To address this issue, we added a pixel-level quantitative comparison between GlobalRice20 and the Cropland Data Layer (CDL) produced by the USDA National Agricultural Statistics Service (USDA NASS). The CDL is an authoritative crop-type reference product for the United States. Specifically, we extracted the rice class from the CDL for 2015 and 2024 and compared it with GlobalRice20 in the major rice-producing regions of the United States. We calculated the intersection over union (IoU), user's accuracy (UA), producer's accuracy (PA), and F1 score for the rice class (Table R11). The results are presented in the table below. GlobalRice20 achieved IoU values of 0.8087 and 0.8319 in 2015 and 2024, respectively, and F1 scores of 0.8942 and 0.9082, respectively.

Table R11. Quantitative spatial agreement between GlobalRice20 and USDA CDL over

the major rice-producing states of the United States.

| Year | Reference | IoU | UA | PA | F1 |
|---|---|---|---|---|---|
| 2015 | USDA CDL | 0.8087 | 0.9162 | 0.8732 | 0.8942 |
| 2024 | USDA CDL | 0.8319 | 0.8961 | 0.9206 | 0.9082 |

These results indicate strong spatial agreement between GlobalRice20 and the CDL. Considering space limitations, we added the IoU metric to Section 4.3.2, "Spatial Consistency with Existing Products," as this metric provides an intuitive measure of the agreement between our results and the CDL. Once again, we sincerely appreciate your valuable suggestion.

The specific revisions are as follows.

**Page 15**

To comprehensively evaluate model performance, we adopted the following four metrics: Overall Accuracy (OA), Producer's Accuracy for the rice class (PA), User's Accuracy for the rice class (UA) and the F1 Score. Additionally, we employed the Intersection over Union (IoU) metric to quantitatively assess the consistency between our product and existing products. Their specific formulas are as follows:

$$\begin{aligned} \mathrm{OA} &= \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \\ \mathrm{PA} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \\ \mathrm{UA} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \\ \mathrm{F1} &= 2 \times \frac{\mathrm{PA}_{\mathrm{rice}} \times \mathrm{UA}_{\mathrm{rice}}}{\mathrm{PA}_{\mathrm{rice}} + \mathrm{UA}_{\mathrm{rice}}} \\ \mathrm{IoU} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \end{aligned} \tag{11}$$

Where TP, TN, FP, and FN represent True Positives, True Negatives, False Positives, and False Negatives, respectively.

**Page 23**

To further ensure the reliability of the comparison, we conducted both qualitative and quantitative comparisons between GlobalRice20 and the USDA National Agricultural Statistics Service (USDA NASS) Cropland Data Layer (CDL; https://croplandcros.scinet.usda.gov/) across the major rice-growing regions of the United States in 2015 and 2024. The CDL provides comprehensive coverage of rice cultivation in the United States and is produced through a well-established and consistent data-processing framework, making it a reliable reference dataset for evaluating regional rice distribution. Visual comparisons show that GlobalRice20 effectively reproduces the major rice-growing zones and field-boundary patterns depicted by the CDL. For the quantitative assessment, the rice class in the CDL was used as the reference layer, and the IoU metric was employed to evaluate pixel-level spatial agreement. The results show that GlobalRice20 achieved IoU scores of 0.8087 and 0.8319 in 2015 and 2024, respectively. Taken together, the quantitative evaluation and visual comparison demonstrate a high degree of spatial consistency between GlobalRice20 and the CDL across the major rice-producing regions of the United States, further confirming the robustness of our dataset across different regions and data standards.

**8. Reporting that 84%/89% of countries exceed OA 0.85 is encouraging, but the remaining countries are not named and the reasons are not given. Please list them with causes, especially where low UA likely explains area overestimation, as flagged for India in 2015.**

**RESPONSE:** We sincerely thank the reviewer for this constructive comment. We agree that simply reporting that 84% of countries in 2015 and 89% of countries in 2024 achieved an OA above 0.85 is not sufficient to fully characterize country-level accuracy differences and uncertainties in GlobalRice20. Therefore, following the reviewer's suggestion, we further examined the country-level accuracy results and provided a complete list of countries with OA values below 0.85 in the response letter. In the

revised manuscript, due to space limitations, we highlight several representative countries and provide additional discussion of the main factors contributing to their relatively lower accuracies.

Specifically, 16 countries had OA values below 0.85 in 2015: Kyrgyzstan, Hungary, Bangladesh, Panama, Indonesia, Bulgaria, Australia, Sierra Leone, Haiti, Guinea, Nicaragua, Thailand, Guyana, Suriname, Burundi, and Costa Rica. In 2024, 11 countries had OA values below 0.85: Nicaragua, Turkmenistan, the Philippines, Brunei, Kyrgyzstan, Australia, Nepal, Sri Lanka, Hungary, South Sudan, and Suriname. The lower accuracies in these countries are likely related to several factors. In some countries, rice fields occupy relatively small areas and are spatially fragmented, meaning that even a limited number of misclassified pixels can substantially affect country-level accuracy. Some tropical and monsoon-region countries are affected by persistent cloud contamination, which limits the availability of valid optical observations. Other regions are characterized by complex cropping calendars, multi-season rice systems, complex terrain, or island landscapes. In addition, spectral and temporal similarities between rice paddies and wetlands, small water bodies, floodplains, or other irrigated crops may also lead to classification confusion. For example, countries such as Kyrgyzstan, Hungary, Bulgaria, Australia, Brunei, Turkmenistan, Guyana, and Suriname have relatively limited rice areas or more fragmented spatial distributions, making their country-level OA more sensitive to local classification errors. Countries such as Bangladesh, Indonesia, Thailand, the Philippines, Nepal, and Sri Lanka are more likely to be affected by persistent cloud contamination, complex cropping systems, field fragmentation, and terrain or island-landscape heterogeneity. The lower accuracies in Sierra Leone, Guinea, South Sudan, Burundi, Haiti, Nicaragua, Panama, and Costa Rica may be associated with tropical cloudy and rainy conditions, insufficient valid observations, confusion between rice paddies and wetlands or water bodies, and insufficiently representative training samples. In addition, lower UA values generally indicate more pronounced commission errors, meaning that some non-rice pixels were incorrectly classified as rice. This may further

lead to overestimation of rice area. This error mechanism is consistent with the overestimation of rice area in India in 2015. Specifically, the optical component in 2015 mainly relied on Landsat-8 imagery, which has a relatively long revisit interval and a native spatial resolution of 30 m, while Sentinel-1 SAR observations were also relatively limited during that period. These factors may have weakened the completeness of temporal feature representation in some regions and reduced the separability of fragmented field boundaries, thereby increasing the likelihood of commission errors and ultimately affecting the stability of area estimates.

We have revised Section 4.3.1 accordingly to improve the transparency of the country-level accuracy assessment and to provide a more complete explanation of accuracy differences among countries and their possible causes. Once again, we sincerely appreciate your valuable suggestion.

The specific revisions are as follows.

**Page 20**

To further characterize country-level uncertainty, we also examined countries with OA values below 0.85. Representative cases included Kyrgyzstan, Hungary, Bangladesh, Indonesia, Bulgaria, and Thailand in 2015, and Nicaragua, the Philippines, Brunei, Kyrgyzstan, Nepal, and Sri Lanka in 2024. The relatively lower accuracies in these countries were mainly associated with limited and fragmented rice cultivation areas, persistent cloud contamination, complex cropping calendars, heterogeneous terrain or island landscapes, and potential confusion between paddy rice and wetlands, small water bodies, floodplains, or other irrigated crops.

Despite these country-level uncertainties, most major rice-producing regions still showed robust classification performance.

**9. It is unclear whether inference used one combined model or two year-specific models. If separate, year-to-year comparability depends on model consistency; if combined, the mixed S2/L8 inputs revisit the cross-sensor issue above.**

**RESPONSE:** Thank you very much for this valuable suggestion. We apologize that the original manuscript did not clearly describe the training and inference strategies used for 2015 and 2024. We would like to clarify that separate T2VRCM models were independently trained and applied for inference in 2015 and 2024, rather than using a single combined model to process data from both years. Specifically, the 2015 model used Landsat-8 and Sentinel-1 data as inputs, whereas the 2024 model used Sentinel-2 and Sentinel-1 data.

Although separate models were trained for the two years, we kept the experimental settings consistent as much as possible to facilitate year-to-year comparison. Specifically, the 2015 and 2024 experiments used stable reference samples from the same locations and the same training/test split, together with the same T2VRCM architecture, input image size, loss function, optimizer, learning rate, number of training epochs, and other training settings. This design minimizes the influence of differences in sample partitioning, model architecture, and training configuration, allowing the two annual products to be evaluated under a consistent experimental framework.

Under these unified experimental settings, the classification performance of the full Optical+SAR models for the two years on the same independent test samples is summarized below.

**Table R12. Classification performance of the year-specific T2VRCM models in 2015 and 2024.**

| Year | Setting | UA (%) | PA (%) | F1 (%) | OA (%) |
|---|---|---|---|---|---|
| 2015 | L8 + S1 | 90.11 | 90.37 | 90.24 | 90.23 |
| 2024 | S2 + S1 | 92.52 | 92.12 | 92.32 | 92.33 |

The results show that, under consistent sample locations and partitioning, model architecture, and training settings, the F1 values for 2015 and 2024 were 90.24% and 92.32%, respectively, while the corresponding OA values were 90.23% and 92.33%. The higher classification performance in 2024 is mainly associated with differences in the actual remote-sensing observation conditions between the two years. In 2015, Landsat-8 was used as the optical data source, with relatively lower temporal resolution, while Sentinel-1 coverage was also comparatively limited. In contrast, the 2024 dataset used Sentinel-2 and benefited from more complete multi-source temporal observations, providing a more comprehensive representation of annual rice phenological dynamics. The cross-sensor differences arising from the use of different optical sensors in the two years have been further discussed and analyzed in our response to the preceding comment.

Following your suggestion, we have further clarified in the Methods section that separate year-specific models were independently trained and applied for inference in 2015 and 2024, thereby improving the clarity of the methodological description. We sincerely thank the reviewer again for this valuable suggestion. The specific revision is provided below.

**Page 16**

Considering the differences in spectral characteristics and temporal sampling between Landsat-8 and Sentinel-2, separate year-specific T2VRCM models were trained for 2015 and 2024 rather than directly applying a single model across the two sensor configurations.

**10. The 20 m label is not uniform: 2024 uses Sentinel-2/Sentinel-1, but 2015 relies on 30 m Landsat-8. Upsampling to 20 m cannot recover sub-30 m detail, so the 2015 layer is effectively ~30 m. The manuscript should state the resampling method and acknowledge this inter-year resolution gap.**

**RESPONSE:** We sincerely thank the reviewer for raising this important point. We agree with the reviewer's observation that resampling 30 m Landsat-8 pixels to a 20 m grid does not create new fine-scale spatial information. We would like to clarify that the term "20 m resolution" in this study refers to the unified spatial grid of the final GlobalRice20 product, rather than implying that all input remote-sensing datasets in both years had identical native spatial resolutions. During data preprocessing, the feature variables derived from the 2015 Landsat-8 OLI optical data were resampled to the 20 m spatial grid using bilinear interpolation to ensure spatial consistency with the 2024 Sentinel-2/Sentinel-1 data and to meet the requirements of subsequent multi-source remote-sensing data fusion and model input. We have added this information to Section 2.2 of the manuscript. This resampling step was used only for spatial grid alignment and does not increase the spatial information content of the original 30 m Landsat-8 data. It should also be emphasized that the 2015 GlobalRice20 product was not generated solely from the resampled Landsat-8 optical data. Instead, it integrated Landsat-8 optical indices with Sentinel-1 VV/VH SAR time-series information. Sentinel-1 SAR data provide finer spatial sampling and all-weather observation capability, contributing important spatial and phenological information for rice mapping in 2015. Therefore, releasing the final product on a unified 20 m grid facilitates consistent use of the dataset across years. We appreciate the reviewer's suggestion regarding differences in the effective scale of spatial information between years. We have added a discussion in Section 4.3.3 on the potential influence of this factor on the slightly lower accuracy of the 2015 product compared with the 2024 product. Once again, we sincerely appreciate your valuable suggestion.

The specific revisions are as follows.

**Page 5**

For 2024, we utilized Sentinel-2 MSI Level-2A surface reflectance data (10-m resolution), applying a cloud-masking algorithm to remove invalid observations. The valid Sentinel-2 observations were then composited at a 5-day interval. For the 2015 target year, as Sentinel-2 data were not yet available, we utilized Landsat-8 OLI surface

reflectance data with a 30-m spatial resolution. After cloud masking, the valid Landsat-8 observations were composited at a 16-day interval.

**Page 25**

The slightly lower accuracy in 2015 can mainly be attributed to the longer revisit cycle and 30 m spatial resolution of Landsat-8 optical imagery, together with the relatively limited SAR coverage in that year. These factors weakened the representation of time-series image features in some regions and may have reduced the delineation of fine field boundaries compared with the 2024 Sentinel-2-based product, thereby affecting the stability and accuracy of area estimation, as reflected in the overestimation of rice area in India in 2015.

**11. Line 211, 224. The format of the parentheses seems incorrect.?**

**RESPONSE:** We apologize for this formatting oversight. The parenthesis formatting and spacing in the indicated lines have been corrected. Meanwhile, we also found similar issues in the parentheses used for equation numbering, and therefore checked and corrected them throughout the manuscript. Once again, we sincerely appreciate your valuable suggestion.

The specific revisions are as follows:

**Pages 11-12**

3.3.1 Overall Architecture

Given an input visual representation $X \in R^{H\times W\times C}$ $(H = W = 224,\ C = 3)$, its forward propagation process is as follows. X first passes through a Stem layer (composed of two 3×3 convolutions with stride 2), performing 4×downsampling to reduce the resolution to $\frac{H}{4}\times\frac{W}{4}$. The output of the Stem, $S(X)$, is then fed into $N_1$ stacked Gate Convolution Blocks $(G_1)$, producing the Stage 1 output feature $O_1 = G_1(S(X))$. Each subsequent

stage i ($i \in \{2,3,4\}$) consists of a downsampling layer ($D_i$, a 3×3 convolution with stride 2) and $N_i$ stacked $G_i$ blocks, yielding $O_i = G_i\left(D_i(O_{i-1})\right)$.

# Response to Dr. Ran Huang

**Comments to the Author**

**This study developed the GlobalRice20 dataset with 20m resolution for 2015 and 2024, constructed a Time-Series-to-Vision framework (T2VRCM) to address cloud contamination and irregular time-series issues in satellite data, and produced a high-accuracy global paddy rice distribution product, which is highly important and practically significant for global food security assessment, agricultural monitoring. The detailed comments are as follows.**

**RESPONSE:** We sincerely thank you for your positive assessment of our work and for your valuable comments and suggestions. Your thoughtful feedback has been highly helpful in improving the clarity, rigor, and overall quality of the manuscript. We have carefully considered each of your comments and provided detailed point-by-point responses below.

**1. Line 204, please discuss whether it is appropriate to directly connect valid observations, particularly when data gaps due to cloud cover are relatively long.**

**RESPONSE:** Thank you very much for this valuable comment. We agree that the treatment of directly connecting valid observations on both sides of relatively long data gaps caused by cloud cover warrants further clarification. We have therefore provided additional explanation and analysis as follows.

First, directly connecting adjacent valid observations in T2V does not imply that the missing interval is treated as a period of continuous observation. Rather, the connecting line is used to describe the overall change between neighboring valid observations, thereby preserving the annual-scale curve shape and phenological variation patterns as much as possible when partial temporal observations are missing. For the optical subplot, the horizontal axis is defined over a unified DOY range of 1–365, and each valid observation is mapped to its corresponding position according to its actual DOY.

The vertical axis represents the actual EVI and LSWI values and is fixed to the range −1.0 to 1.0. When a long temporal gap occurs between two valid observations, no additional observation markers are introduced within the missing interval; the connecting line only links the valid observations that actually exist on the two sides of the gap, thereby preserving the overall annual pattern of optical-index variation.

In addition, to further evaluate the suitability of this direct-connection strategy, we conducted a controlled comparison of temporal-gap representations using the 2024 dataset. The EVI, LSWI, VV, and VH input variables, DOY coordinate range, image size, T2VRCM architecture, training and test samples, and training parameters were kept identical, and only the treatment of missing intervals was changed. The original Current T2V representation connects adjacent valid observations according to their actual DOY order, even when missing intervals occur between them. As a comparison, the Gap-explicit T2V representation connects only temporally continuous valid observations; when a missing interval occurs between two adjacent valid observations, the curve is explicitly disconnected and the corresponding temporal interval is left blank.

**Table R13. Comparison between the current T2V representation and the gap-explicit temporal-availability representation.**

| Representation | UA (%) | PA (%) | F1 (%) | OA (%) |
|---|---|---|---|---|
| Current T2V | 92.52 | 92.12 | 92.32 | 92.33 |
| Gap-explicit T2V | 91.58 | 91.32 | 91.45 | 91.46 |

The experimental results show that, under the current dataset and experimental settings, explicitly disconnecting the curves across missing intervals did not improve classification accuracy. Compared with Gap-explicit T2V, Current T2V increased F1 from 91.45% to 92.32%, corresponding to an improvement of 0.87 percentage points, while OA increased from 91.46% to 92.33%. These results indicate that retaining connections between valid observations on a unified DOY time axis is more favorable for rice classification in the current task. This is mainly because directly connecting valid observations better preserves the overall annual curve shape and the temporal

variation characteristics associated with rice growth, whereas explicitly disconnecting the curves across missing intervals increases visual fragmentation and weakens the overall representation of long-term phenological variation. In addition, when cloud cover causes long temporal gaps and valid EVI and LSWI observations become sparse, the amount of phenological information available from the optical time series is correspondingly reduced. In such cases, T2VRCM can still utilize the temporal backscatter information provided by Sentinel-1 VV and VH as complementary information, thereby reducing the dependence of the classification process on the completeness of the optical time series.

Overall, the current T2V representation preserves the annual-scale phenological curve structure by connecting valid observations on both sides of missing intervals, while Sentinel-1 SAR time-series features provide complementary information when optical observations become sparse because of persistent cloud cover. The additional experiment further demonstrates that, compared with explicitly disconnecting missing intervals, the current T2V representation with direct connections achieves higher overall classification accuracy, indicating that this treatment is well suited to the task and experimental settings considered in this study. We sincerely thank the reviewer again for this valuable suggestion.

**2. The middle sections of Figure 3 and Figure 4 are mostly duplicated. Please adjust the figures.**

**RESPONSE:** Thank you very much for this valuable suggestion. Following your comment, we revised Figures 3 and 4 to reduce redundant content between them and to further improve the clarity and overall visual presentation of the figures. We sincerely appreciate your helpful suggestion.

The corresponding revisions are shown below.

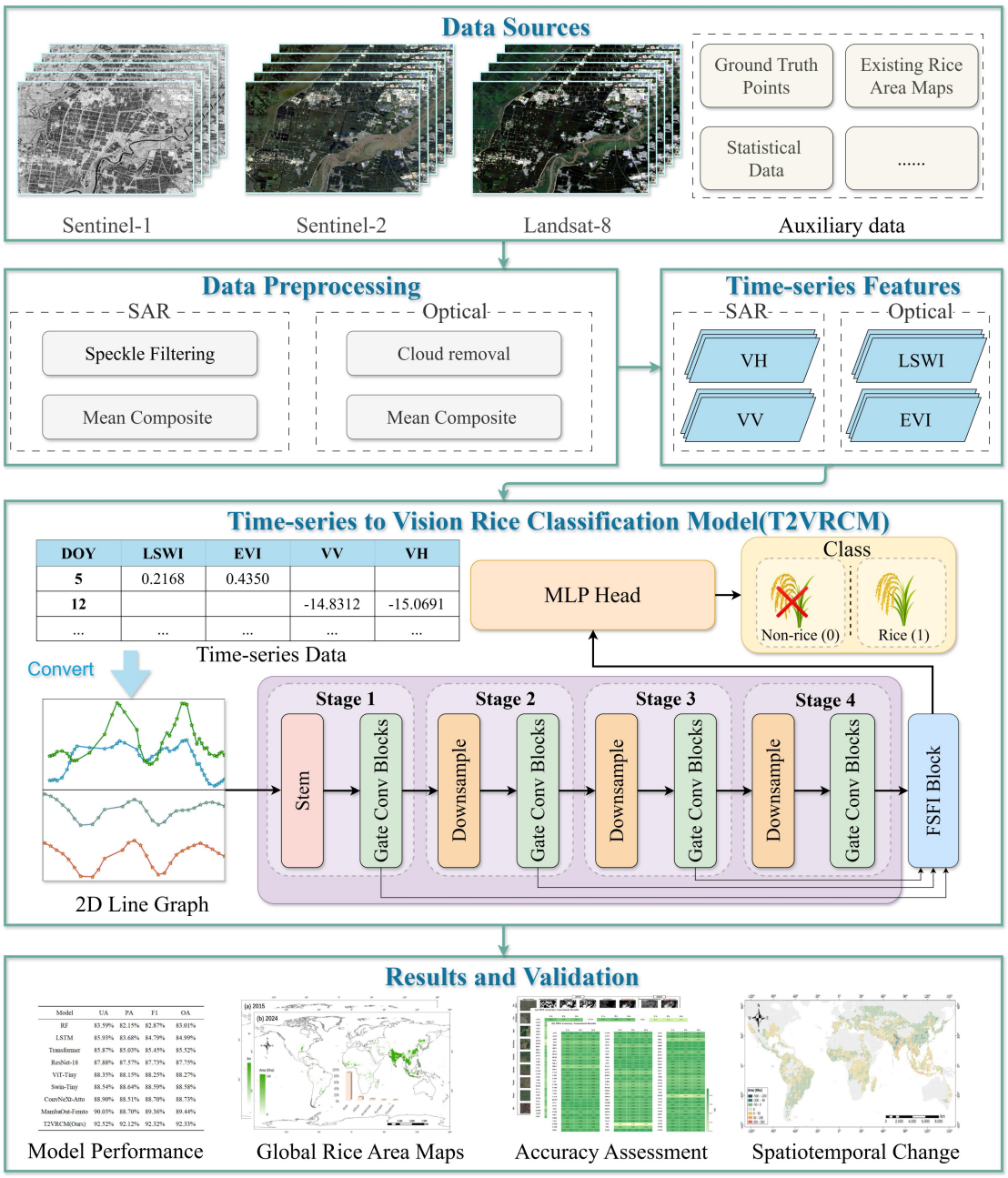


**Figure 3. Technical Workflow**

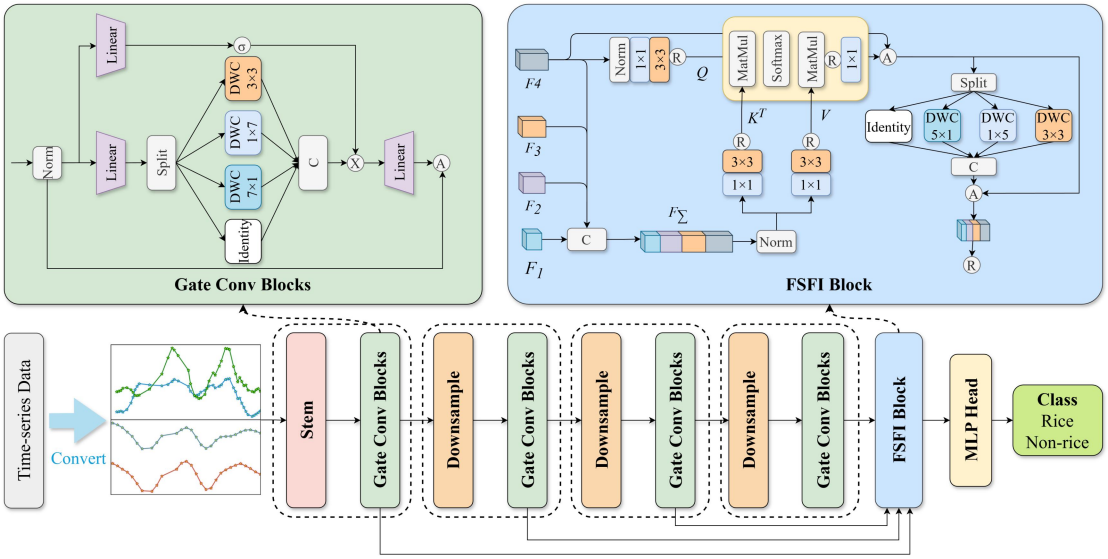


**Figure 4. Overall Architecture of T2VRCM**

**3. Rice growing periods differ greatly between the Northern Hemisphere, Southern Hemisphere, and equatorial zone. Please specify how these differences were accounted for during model training.**

**RESPONSE:** Thank you very much for this valuable comment. We agree that, because of differences in climatic conditions and cropping systems, the timing of rice sowing, growth, and harvesting varies substantially across the Northern Hemisphere, Southern Hemisphere, and equatorial regions. To account for these global differences in rice phenology, we considered three main aspects: sample representativeness, the training/test sample partitioning strategy, and full-year temporal representation.

First, the reference samples used in this study cover 98 major rice-producing countries across six continents, encompassing a wide range of latitudes, climatic conditions, and cropping systems, including regions dominated by single- and double-cropping rice. This broad spatial coverage enables the training samples to represent the major rice phenological patterns occurring globally. Second, during the training/test split, samples from each country were divided into training and test subsets at a ratio of 80%:20%. The training samples from all countries for the corresponding year were then

combined to train the year-specific global T2VRCM model, while the independently retained test samples from each country were used for evaluation. This strategy ensures that samples from different countries, latitudes, and cropping systems are all represented during model training, thereby reducing the risk of the training data being overly concentrated in a particular region or rice-growing period.

In addition, T2VRCM does not rely on a predefined common rice-growing season or a fixed phenological window, nor do we artificially shift or phenologically align the time series from different hemispheres. For all samples, the complete annual time series of EVI, LSWI, VV, and VH are retained and represented according to their actual Day of Year (DOY) positions. In this way, the actual timing of rice growth and phenological changes in different regions is preserved in the T2V images.

Therefore, through broad multi-regional sample coverage, country-level training/test partitioning, and full-year DOY-based temporal representation, the model is able to learn regional differences in rice growing periods, phenological dynamics, and cropping cycles during training. This allows T2VRCM to accommodate the spatial heterogeneity of rice phenology across the Northern Hemisphere, Southern Hemisphere, and equatorial regions without requiring manually specified region-specific or hemisphere-specific rice calendars. We sincerely thank the reviewer again for this valuable suggestion.

**4. In Section 4.2, please analyze the model accuracy obtained using only Sentinel-2 or Landsat-8 data in the results section, and quantify how much the accuracy is improved when adding Sentinel-1 data.**

**RESPONSE:** Thank you very much for this valuable suggestion. We agree that, although the original manuscript reported the classification performance of the full T2VRCM model, it did not separately analyze the performance obtained using only Landsat-8 or Sentinel-2 optical data, nor did it quantitatively evaluate the accuracy improvement achieved by incorporating Sentinel-1. To address this issue, we conducted

additional modality ablation experiments using the 2015 and 2024 datasets, respectively. Three input configurations were considered: Optical-only, SAR-only, and Optical+SAR. The sample split, T2VRCM architecture, and training settings were kept consistent across all experiments to quantitatively assess the contribution of different data modalities to rice classification performance.

**Table R14. Classification performance of T2VRCM under different modality configurations in 2015 and 2024.**

| Year | Setting | UA (%) | PA (%) | F1 (%) | OA (%) |
|---|---|---|---|---|---|
| 2015 | Optical+SAR (L8 + S1) | 90.11 | 90.37 | 90.24 | 90.23 |
| 2015 | Optical-only (L8) | 84.93 | 86.00 | 85.46 | 85.37 |
| 2015 | SAR-only (S1) | 81.19 | 80.91 | 81.05 | 81.08 |
| 2024 | Optical+SAR (S2 + S1) | 92.52 | 92.12 | 92.32 | 92.33 |
| 2024 | Optical-only (S2) | 86.76 | 87.62 | 87.19 | 87.13 |
| 2024 | SAR-only (S1) | 85.80 | 86.54 | 86.17 | 86.11 |

The experimental results show that, when only optical data were used, the F1 values were 85.46% for Landsat-8 in 2015 and 87.19% for Sentinel-2 in 2024, while the corresponding OA values were 85.37% and 87.13%, respectively. After incorporating Sentinel-1, the 2015 F1 increased to 90.24%, representing an improvement of 4.78 percentage points over the Optical-only setting, while OA increased from 85.37% to 90.23%, corresponding to an improvement of 4.86 percentage points. For 2024, the addition of Sentinel-1 increased F1 from 87.19% to 92.32%, an improvement of 5.13 percentage points, while OA increased from 87.13% to 92.33%, corresponding to an improvement of 5.20 percentage points.

These results demonstrate that, regardless of whether Landsat-8 or Sentinel-2 was used as the optical data source, incorporating Sentinel-1 SAR time-series observations consistently improved rice classification performance. This finding indicates clear complementarity between optical temporal information and SAR backscatter dynamics

for rice identification and further demonstrates the effectiveness of multi-source data fusion for global rice classification. We sincerely thank the reviewer again for this valuable suggestion.

**5. In Section 4.3.1 please clarify how the 164,000 global samples were split into training, validation, and test sets.**

**RESPONSE:** Thank you for your valuable feedback. Regarding the specific partitioning method for the 164,000 global reference samples, we have provided a detailed explanation in Section 2.3.1 (Reference Sample Set) of the revised manuscript. We believe that the sample partitioning methodology inherently belongs to the dataset construction process. Therefore, placing it in the "Materials" section (Section 2.3.1) rather than the "Experimental Results" section (Section 4.3.1) better aligns with the overall logical structure of the paper. The specific partitioning scheme for the sample set is as follows:

First, the 164,000 samples were spatially stratified using countries as the spatial stratification units. Within each country, spatially stratified random partitioning was conducted, maintaining a fixed 4:1 ratio for the training and test sets (yielding approximately 131,200 training samples and 32,800 test samples globally). Second, a 1:1 class balance between rice and non-rice samples was preserved in both the training and test sets.

Furthermore, it should be noted that this study did not partition a separate, independent validation set. As described in Section 3.5 (Model Comparison) of the original manuscript, all deep learning models were trained using a unified configuration: the AdamW optimizer with an initial learning rate of 1e-3 combined with a cosine learning rate decay schedule, a batch size of 64, and a fixed training duration of 100 epochs. We did not employ an early stopping mechanism based on a validation set, nor did we conduct hyperparameter searches relying on a validation set. The primary intention behind this design was to ensure that conditions such as the number of training epochs and optimizer settings remained completely unified and fairly comparable when

evaluating T2VRCM against the baseline models. Consequently, partitioning the data into only two subsets—training and test sets—was sufficient to meet our experimental design requirements, eliminating the need for an additional, independent validation set.

The aforementioned explanations regarding the training/test set split ratio, spatial stratification approach, and the handling of the validation set have now been presented as a complete paragraph in Section 2.3.1, mutually corroborating the training configuration described in Section 3.5. We believe this clarification fully addresses your query regarding the sample partitioning method mentioned in Section 4.3.1. Thank you once again for helping us further clarify the details of our experimental design.

The specific revisions are as follows.

**Page 7**

The sample points were partitioned into training and test sets employing a spatially stratified random sampling strategy, utilizing each country as a stratification unit. Within each country, samples were allocated to the training and test sets at a fixed ratio of 4:1 (yielding approximately 131,200 training samples and 32,800 test samples globally). A balanced 1:1 ratio between rice and non-rice samples was maintained in both subsets to prevent class imbalance.

Crucially, to ensure temporal consistency and reduce label noise, we adopted a "stable sample" strategy: selected samples were restricted to locations that maintained a consistent land cover type (either persistent rice or persistent non-rice) in both 2015 and 2024. Consequently, identical sample locations and training/test splits were applied for both 2015 and 2024. However, we independently trained and evaluated two separate T2VRCM models for the two target years, utilizing the time-series features corresponding to each respective year. The training and evaluation of each model were strictly confined to the contemporary training and test sets of that specific year.

**6. In line 409, the statement "reveal a high spatial agreement" is qualitative. Please provide quantitative metrics to specify the level of agreement, rather than using a qualitative description.**

**RESPONSE:** Thank you very much for this insightful comment. We agree that the original phrase "reveal a high spatial agreement" was qualitative and did not provide sufficient quantitative evidence. To address this, we conducted a pixel-level quantitative comparison between GlobalRice20 and USDA CDL over the major rice-producing states in the United States. Specifically, we calculated IoU, UA, PA, and F1 scores for the rice class. The results show that GlobalRice20 achieved IoU values of 0.8087 in 2015 and 0.8319 in 2024, with corresponding UA values of 0.9162 and 0.8961, PA values of 0.8732 and 0.9206, and F1 scores of 0.8942 and 0.9082, respectively. These results provide quantitative evidence of the strong spatial agreement between GlobalRice20 and CDL. In the revised manuscript, we added the IoU metric in Section 4.3.2 (Spatial Consistency with Existing Products) and revised the qualitative description accordingly. Once again, we sincerely appreciate your valuable suggestion.

The specific revisions are as follows.

**Page 15**

To comprehensively evaluate model performance, we adopted the following four metrics: Overall Accuracy (OA), Producer's Accuracy for the rice class (PA), User's Accuracy for the rice class (UA) and the F1 Score. Additionally, we employed the Intersection over Union (IoU) metric to quantitatively assess the consistency between our product and existing products. Their specific formulas are as follows:

$$OA = \frac{TP + TN}{TP + TN + FP + FN}$$

$$PA = \frac{TP}{TP + FN}$$

$$UA = \frac{TP}{TP + FP} \quad (11)$$

$$F1 = 2 \times \frac{PA_{rice} \times UA_{rice}}{PA_{rice} + UA_{rice}}$$

$$IoU = \frac{TP}{TP + FP + FN}$$

Where TP, TN, FP, and FN represent True Positives, True Negatives, False Positives, and False Negatives, respectively.

**Page 23**

To further ensure the reliability of the comparison, we conducted both qualitative and quantitative comparisons between GlobalRice20 and the USDA National Agricultural Statistics Service (USDA NASS) Cropland Data Layer (CDL; https://croplandcros.scinet.usda.gov/) across the major rice-growing regions of the United States in 2015 and 2024. The CDL provides comprehensive coverage of rice cultivation in the United States and is produced through a well-established and consistent data-processing framework, making it a reliable reference dataset for evaluating regional rice distribution. Visual comparisons show that GlobalRice20 effectively reproduces the major rice-growing zones and field-boundary patterns depicted by the CDL. For the quantitative assessment, the rice class in the CDL was used as the reference layer, and the IoU metric was employed to evaluate pixel-level spatial agreement. The results show that GlobalRice20 achieved IoU scores of 0.8087 and 0.8319 in 2015 and 2024, respectively. Taken together, the quantitative evaluation and visual comparison demonstrate a high degree of spatial consistency between GlobalRice20 and the CDL across the major rice-producing regions of the United States, further confirming the robustness of our dataset across different regions and data standards.

# Response to Dr. Dailiang Peng

**Comments to the Author**

**GlobalRice20 gives us the first validated, globally consistent 20 m paddy rice map for 2015 and 2024, built from multi-source optical and SAR data through the Time-Series-to-Vision framework. It is a timely and genuinely useful contribution to food-security and SDG 2 work. However, before finalizing the published paper, I believe that some of its content still requires further clarification or modification.**

**RESPONSE:** We sincerely appreciate your encouraging evaluation of the GlobalRice20 dataset and the T2VRCM framework, as well as your constructive suggestions for further improving the manuscript. Your comments have provided valuable guidance for enhancing the scientific presentation, methodological clarity, and overall readability of our work. We have carefully considered all of your suggestions and addressed them individually in the responses below.

**1. The manuscript currently mentions the use of 164,000 global samples for the experiments, but the specific partitioning method for these samples is not clearly described. It is recommended to further clarify in the text exactly how these 164,000 global samples were divided into training and test sets.**

**RESPONSE:** Thank you for your valuable feedback. Regarding the specific partitioning method for the 164,000 global samples, we have provided explicit clarifications in Section 2.3.1 (Reference Sample Set) of the revised manuscript. The core details are as follows:

We partitioned the 164,000 samples into training and test sets at a 4:1 ratio, yielding approximately 131,200 training samples and 32,800 test samples globally.

We strictly maintained a 1:1 class balance between rice and non-rice samples in both the training and test sets.

Utilizing each country as a spatial stratification unit, we conducted independent random partitioning within each country. This ensures that different countries are represented in both the training and test sets, and it reduces the risk of accuracy overestimation caused by adjacent or duplicate samples appearing simultaneously in both subsets.

Based on the "stable sample" strategy, both 2015 and 2024 share the exact same geographic locations for the sample points and their corresponding training/test set assignments. However, we independently trained two separate T2VRCM models for the two target years. The training and testing of each model were strictly confined to the contemporary training and test subsets of that respective year, ensuring no cross-year data leakage.

The aforementioned details are now clearly presented as an independent paragraph in Section 2.3.1. We believe this addition fully addresses your concerns regarding the transparency of the sample partitioning method, facilitates readers' understanding of the dataset construction workflow, and provides the essential information necessary for future benchmarking reproduction. Thank you once again for your suggestions, which have significantly helped us refine the descriptions in the methodology section.

The specific revisions are as follows.

**Page 7**

The sample points were partitioned into training and test sets employing a spatially stratified random sampling strategy, utilizing each country as a stratification unit. Within each country, samples were allocated to the training and test sets at a fixed ratio of 4:1 (yielding approximately 131,200 training samples and 32,800 test samples globally). A balanced 1:1 ratio between rice and non-rice samples was maintained in both subsets to prevent class imbalance.

Crucially, to ensure temporal consistency and reduce label noise, we adopted a "stable sample" strategy: selected samples were restricted to locations that maintained a consistent land cover type (either persistent rice or persistent non-rice) in both 2015 and

2024. Consequently, identical sample locations and training/test splits were applied for both 2015 and 2024. However, we independently trained and evaluated two separate T2VRCM models for the two target years, utilizing the time-series features corresponding to each respective year. The training and evaluation of each model were strictly confined to the contemporary training and test sets of that specific year.

**2. In the results analysis of Section 4.2, it is recommended to supplement the model's accuracy performance when utilizing only optical data (e.g., Sentinel-2 in 2024 or Landsat-8 in 2015). Furthermore, please quantitatively analyze the specific magnitude of accuracy improvement brought by the addition of Sentinel-1 radar data, thereby better highlighting the necessity of multi-source data fusion.**

**RESPONSE:** Thank you very much for this valuable suggestion. We agree that reporting only the classification results obtained from multi-source data fusion does not fully demonstrate the contribution of Sentinel-1 SAR data to rice identification. To address this issue, we conducted additional modality ablation experiments using the 2015 and 2024 datasets, respectively. Three input configurations were considered: Optical-only, SAR-only, and Optical+SAR. The sample split, T2VRCM architecture, and training settings were kept consistent across all experiments to quantitatively assess the effects of different data modalities on rice classification performance.

**Table R15. Classification performance of T2VRCM under different modality configurations in 2015 and 2024.**

| Year | Setting | UA (%) | PA (%) | F1 (%) | OA (%) |
|---|---|---|---|---|---|
| 2015 | Optical+SAR (L8 + S1) | 90.11 | 90.37 | 90.24 | 90.23 |
| 2015 | Optical-only (L8) | 84.93 | 86.00 | 85.46 | 85.37 |
| 2015 | SAR-only (S1) | 81.19 | 80.91 | 81.05 | 81.08 |
| 2024 | Optical+SAR (S2 + S1) | 92.52 | 92.12 | 92.32 | 92.33 |

| Year | Setting | UA (%) | PA (%) | F1 (%) | OA (%) |
|---|---|---|---|---|---|
| 2024 | Optical-only (S2) | 86.76 | 87.62 | 87.19 | 87.13 |
| 2024 | SAR-only (S1) | 85.80 | 86.54 | 86.17 | 86.11 |

The experimental results show that, when only optical data were used, the F1 values were 85.46% for Landsat-8 in 2015 and 87.19% for Sentinel-2 in 2024, while the corresponding OA values were 85.37% and 87.13%, respectively. After incorporating Sentinel-1, the 2015 F1 increased to 90.24%, representing an improvement of 4.78 percentage points over the Optical-only setting, while OA increased from 85.37% to 90.23%, corresponding to an improvement of 4.86 percentage points. For 2024, the addition of Sentinel-1 increased F1 from 87.19% to 92.32%, an improvement of 5.13 percentage points, while OA increased from 87.13% to 92.33%, corresponding to an improvement of 5.20 percentage points.

These results demonstrate that, regardless of whether Landsat-8 or Sentinel-2 was used as the optical data source, incorporating Sentinel-1 SAR time-series observations consistently improved classification performance. This finding indicates strong complementarity between optical temporal information and SAR backscatter dynamics for rice identification and further demonstrates the effectiveness of multi-source data fusion for global rice classification. We sincerely thank the reviewer again for this valuable suggestion.

**3. To improve the readability of the tables, it is recommended to highlight the best-performing model results in bold in both Table 1 and Table 2.**

**RESPONSE:** Thank you very much for this valuable suggestion. Following your comment, we have highlighted in bold the best-performing results for each evaluation metric in Tables 1 and 2, allowing readers to more readily identify and compare the classification performance of different models and experimental settings and thereby

improving the readability of the tables. We sincerely thank you again for this helpful suggestion.

**4. In Table 2, which presents the ablation study results, adding an extra column or using a clearer notation (e.g., adding "+X.XX%" in parentheses) to show the specific performance improvement relative to the baseline would help to more intuitively reflect the actual contributions of each module.**

**RESPONSE:** Thank you very much for this valuable suggestion. We agree that explicitly indicating the performance improvement of each module relative to the baseline in the ablation study helps more clearly illustrate the actual contribution of different components. Following your suggestion, we have added the corresponding performance gains relative to the baseline in Table 2 and presented them in the form of "+X.XX%" to facilitate a more intuitive comparison of the improvements introduced by each module. We sincerely thank you again for this helpful suggestion.

**5. In Figure 1, the font size within the legends appears slightly uncoordinated compared to the typography of the rest of the map. It is recommended to appropriately adjust and standardize the font sizes in these legends to further enhance the overall aesthetic appeal of the figure.**

**RESPONSE:** Thank you for your valuable feedback. We fully agree that coordinated and consistent formatting is essential for enhancing the professionalism and aesthetic appeal of the figures. Based on your suggestion, we have made the following modifications to Figure 1:

First, we have uniformly adjusted the font and font size of all text elements across all subplots in Figure 1—including the legend titles, legend items, scale bar labels, north arrow labels, and axis labels—to ensure consistency with the overall typographic style of the figure.

Second, we have redesigned the layout of the legend in subplot (b), optimizing the arrangement order and spacing of the legend items to achieve a more natural, coordinated, and aesthetically pleasing appearance.

We believe that these modifications have effectively enhanced the overall visual consistency and professional presentation of Figure 1. The revised Figure 1 has been updated in the revised manuscript.

**Page 4**

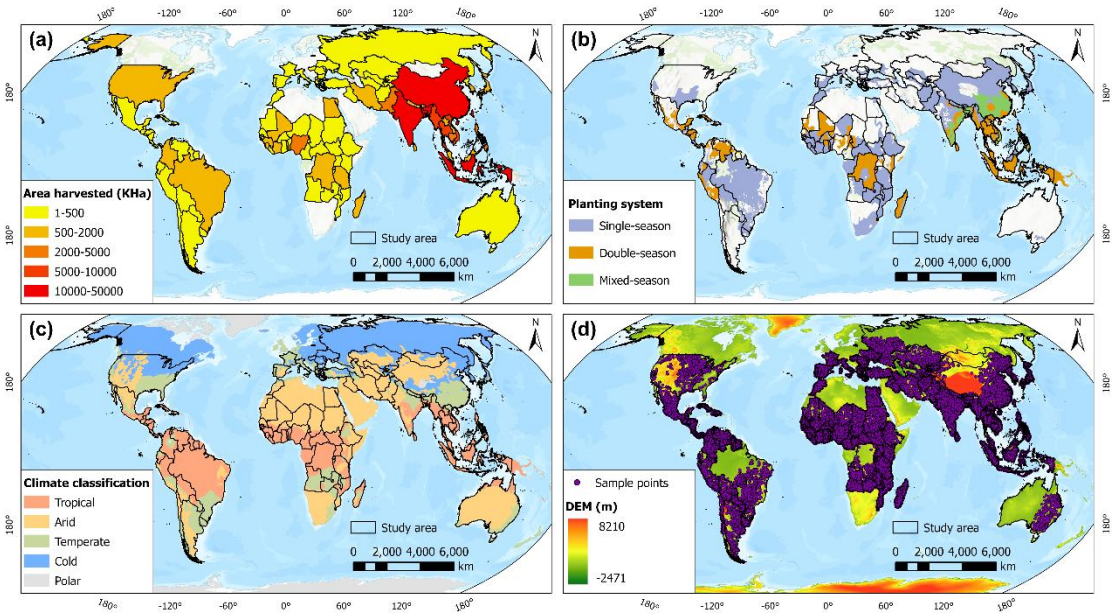


**Figure 1. Overview of the study area. (a) Rice planting areas, (b) Rice cropping patterns, (c) Climate zoning, (d) Distribution of validation sample points.**

**6. There is a certain degree of content overlap between Figure 3 and Figure 4. It is recommended to appropriately simplify the model visualization of the T2VRCM component in Figure 3 to avoid excessive redundancy with Figure 4.**

**RESPONSE:** Thank you very much for this valuable suggestion. Following your comment, we revised and optimized the presentation of Figures 3 and 4 to reduce redundant information between the two figures and to clarify their respective roles in illustrating the overall workflow and the detailed model architecture. These revisions

improve the readability and consistency of the figures. We sincerely thank you again for this helpful suggestion.

The corresponding revisions are shown below.

**Page 9**

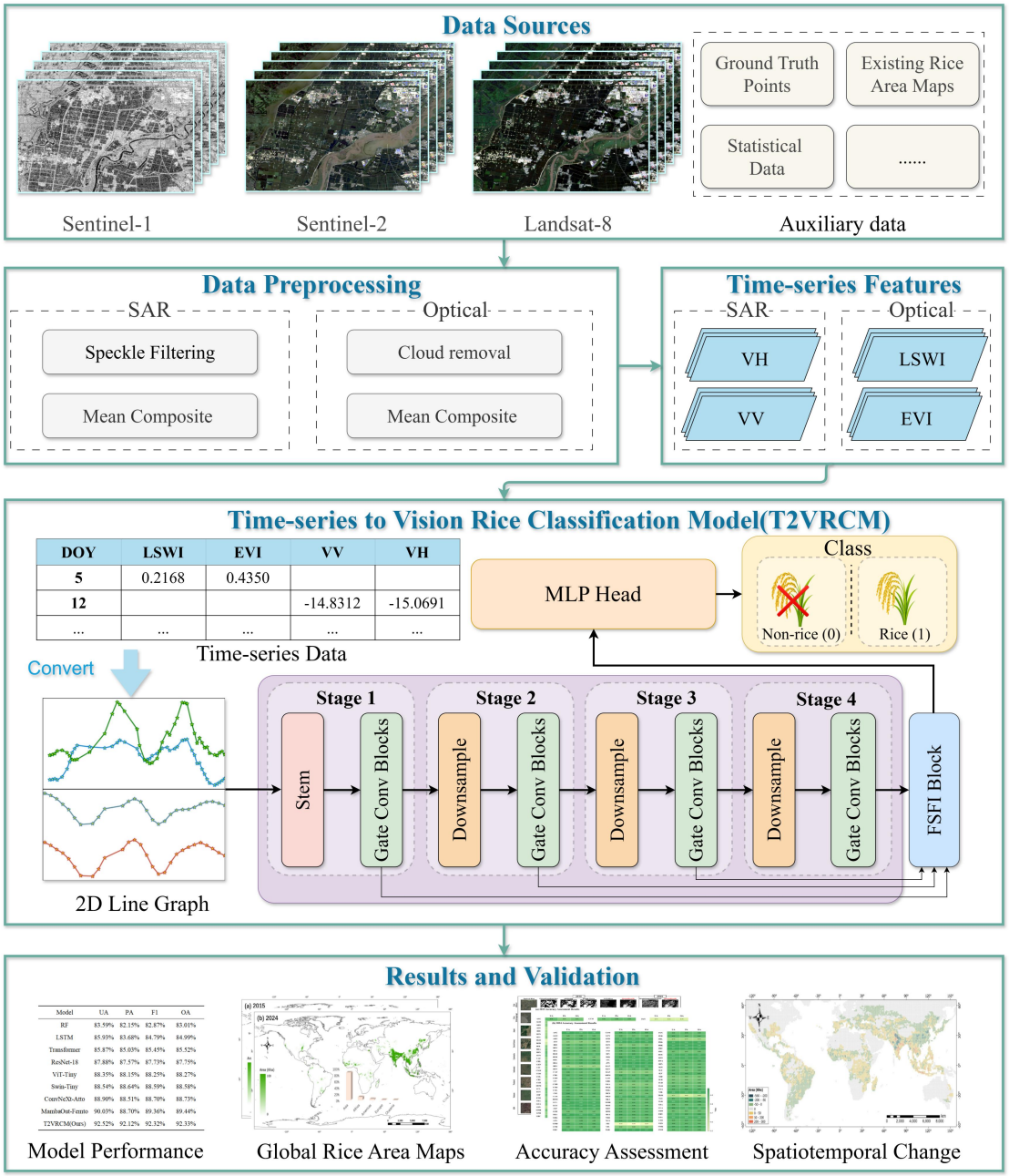


**Figure 3. Technical Workflow**

## Page 12

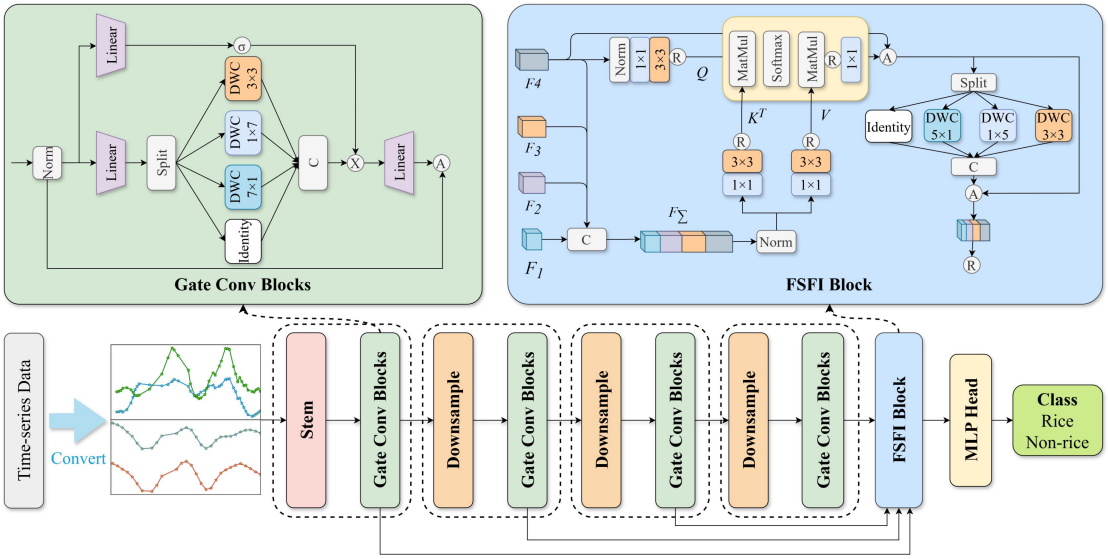


**Figure 4. Overall Architecture of T2VRCM**

**7. The core concept "Time-Series-to-Vision" appears multiple times in the text. However, the capitalization forms (e.g., "Time-series-to-Vision" vs. "Time-Series-to-Vision") are currently used interchangeably and are not fully consistent. It is recommended to conduct a full-text search and unify them into a standard format.**

**RESPONSE:** Thank you very much for this valuable suggestion. Following your comment, we carefully reviewed the relevant terminology throughout the manuscript and standardized all occurrences to “Time-Series-to-Vision” to ensure consistent usage of this core concept throughout the paper. We sincerely appreciate your careful and helpful suggestion.

**8. In Section 3.4, the metrics are denoted as UArice and PArice, while the subsequent Table 2 and Figure 6 use UA and PA. It is recommended to standardize the names of these relevant indicators throughout the manuscript to avoid any ambiguity.**

**RESPONSE:** Thank you very much for this valuable suggestion. Following your comment, we reviewed the relevant evaluation metrics throughout the manuscript and standardized the terminology used in the main text, tables, and figures to UA and PA to ensure consistency and avoid ambiguity. We sincerely thank you again for this helpful suggestion.

**9. The capitalization of "KM" and "km" in the spatial distribution maps is inconsistent. It is recommended to unify their representation (preferably using the standard lowercase "km").**

**RESPONSE:** Thank you for your valuable feedback. We fully agree that uniformly adopting the standard lowercase "km" in the spatial distribution maps is more rigorous and standardized. Following your suggestion, we have thoroughly reviewed all figures involving spatial scale units in the manuscript, with the following results:

The scale bar units in Figure 5 and Figure 10 were already utilizing the standard lowercase "km", which complies with standard conventions and required no modification.

The scale bar units in Figure 1 previously contained the non-standard uppercase "KM". We have now uniformly corrected this to the standard lowercase "km" to maintain consistency with Figure 5 and Figure 10.

Additionally, we have conducted a comprehensive review of the expressions involving distance and area units throughout the main text and figure captions, confirming that there are no other instances of inconsistent capitalization.

Following these revisions and verifications, all scale and distance unit annotations across the spatial distribution maps and the main text have now been completely standardized to "km". Thank you once again for helping us further enhance the formatting rigor and meticulousness of our manuscript.

**10. In the in-text labels of Figure 8, the coefficient of determination is currently written as "R2". It is recommended to format this using the standard mathematical superscript as R2 to comply with rigorous academic publishing standards.**

**RESPONSE:** Thank you very much for your comment. We sincerely apologize for this formatting oversight. Following the reviewer's suggestion, we have revised the in-text labels in Figure 8 and changed "R2" to the standard mathematical notation "$R^2$" to comply with academic publishing standards. Once again, we sincerely appreciate your valuable suggestion.

The specific revisions are as follows:

**Page 25**

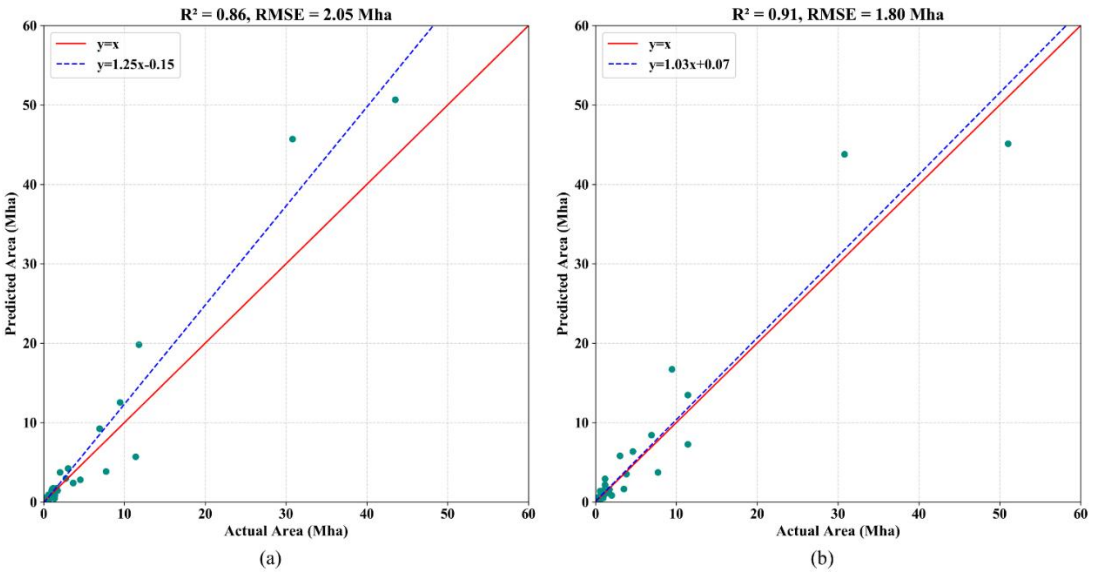


**Figure 8. Accuracy validation based on statistical data. (a) $R^2$ between predicted and statistical rice areas by country in 2015. (b) $R^2$ between predicted and statistical rice areas by country in 2024.**