

Response to the Community and Reviewers

We sincerely appreciate the Community and Reviewers for their constructive comments on our manuscript, “Constructing nationally comprehensive annual rice paddy maps for Madagascar from 2017 to 2025.” Original comments appear in *black italics*, our replies in blue, and revised manuscript passages in green. The responses are organized below for Reviewer 1, Reviewer 2, and the Community.

Summary of the key changes

1. We have replaced “near label-free” with “phenology-guided weak supervision with uncertainty-guided refinement” and clarified the contributions of pseudo-labels and manually interpreted samples.
2. We have clarified how reference samples were labeled, assessed the effect of nearby training samples on accuracy, and made all 2,000 reference samples publicly available.
3. We have strengthened comparisons with Jiang et al. (2025) and De Vos et al. (2023) and revised our explanations of the observed mapping differences.
4. We have clarified that cropping intensity includes rice followed by another crop, so two NDVI peaks do not necessarily indicate two rice cultivation cycles.
5. We have estimated uncertainty in national rice area and net change following Olofsson et al. (2014) and explained the assumptions underlying these estimates.

19 **Reviewer 1**

20 *This study develops a 10-m annual rice paddy mapping framework for Madagascar from 2017 to 2025.*
21 *The method combines phenology-guided pseudo-label generation from HLS time series with Google*
22 *Satellite Embedding (GSE) features and a two-stage random forest classifier, in which manually*
23 *interpreted hard samples are further used to refine low-confidence areas. The study addresses an*
24 *important challenge in large-scale crop mapping under limited training data and provides a potentially*
25 *valuable long-term rice dataset for Madagascar. However, several issues regarding the near-label-free*
26 *claim, training-validation independence, and benchmark comparability should be further clarified*
27 *before the robustness and generalizability of the framework can be fully assessed.*

28 *Response: Thank you very much for the careful assessment and for recognizing the value of the annual*
29 *rice paddy dataset for Madagascar. We have clarified the role of manually interpreted samples, evaluated*
30 *the separation between training and validation, and revised the comparison with the existing Africa-*
31 *wide rice map. We have addressed each concern below.*

32

33 **Major Comment 1.**

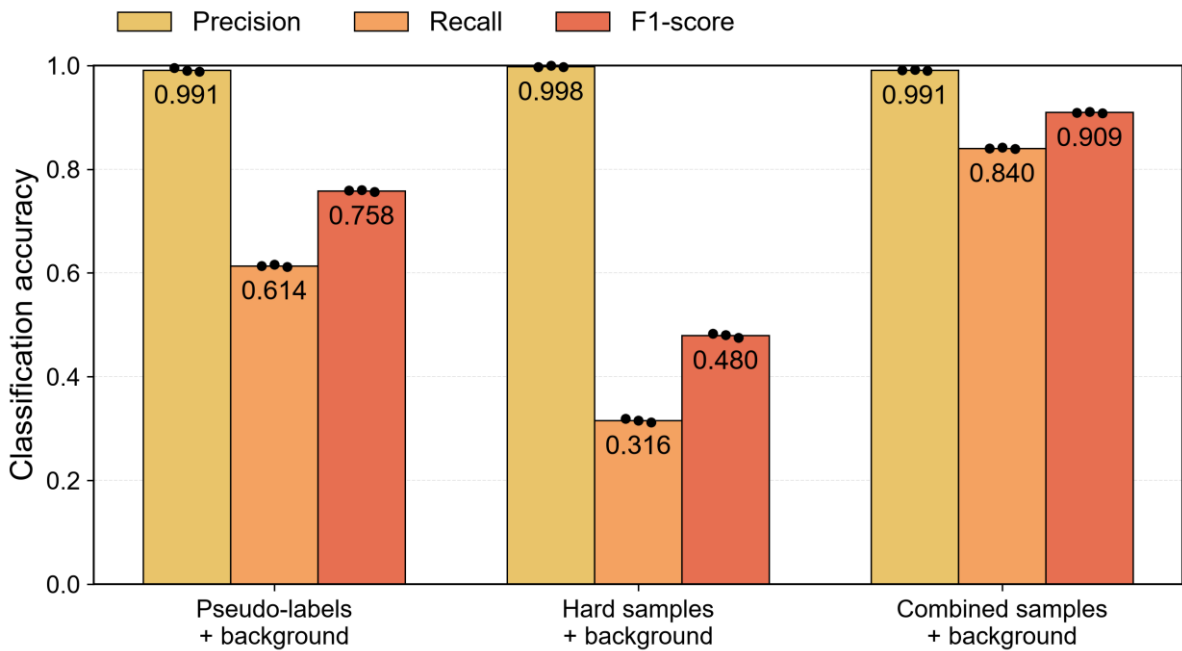
34 *The manuscript repeatedly characterizes the proposed workflow as “near label-free” and emphasizes*
35 *that it avoids costly field reference data. However, the final second-stage classifier is refined by adding*
36 *1,000 manually labeled hard samples per class per year from low-confidence regions. Given three target*
37 *classes and nine years (2017–2025), this design may involve up to 27,000 manually interpreted samples.*
38 *More importantly, the reported F1-score increases markedly from 0.728 for the first-stage pseudo-label-*
39 *only classifier to 0.904 for the final second-stage classifier. Therefore, the strongest reported accuracy*
40 *is not achieved by the near-label-free component alone, but only after substantial targeted manual*
41 *intervention. This raises a fundamental question about whether “near label-free” accurately*
42 *characterizes the final mapping framework and how much of the reported performance can actually be*
43 *attributed to the pseudo-label-based component.*

44 *Response: Thank you very much for highlighting the contribution of manual refinement to the final*
45 *accuracy. We agree that “near label-free” does not fully describe the final framework. We have therefore*
46 *replaced this wording throughout the manuscript with “phenology-guided weak supervision with*
47 *uncertainty-guided refinement.” In the first stage, we applied phenological rules to HLS-derived NDVI*
48 *and LSWI time series to generate pseudo-labels for rice paddies, other crops, and wetlands. We used*
49 *these labels with GSE annual features to train the first-stage classifier. In the second stage, model*
50 *confidence guided the manual interpretation of difficult areas, and samples extracted from these areas*
51 *supplemented the pseudo-labels to refine the classification. This description explicitly acknowledges*
52 *manual refinement as part of the final framework while distinguishing it from the automatic generation*
53 *of pseudo-labels.*

54 *We have clarified the manual labeling process because our original description could have given the*
55 *impression that we independently interpreted 27,000 samples. We selected 36 polygons representing*
56 *difficult rice, other-crop, and wetland cases from low-confidence areas identified in the first-stage*
57 *classification confidence maps, rather than choosing them randomly. These polygons were distributed*
58 *across 19 of Madagascar’s 22 administrative regions and ranged from 15.1 to 934.1 ha, with a median*
59 *area of 138.2 ha. For each year from 2017 to 2025, we reviewed Sentinel-2 RGB imagery and Google*
60 *Earth imagery, where available, to check for crop rotation, fallow periods, and land-use changes before*

61 assigning the polygon labels. We then generated hard samples through stratified random sampling
 62 within these polygons at the 10 m resolution of GSE. The resulting 3,000 sample locations per year
 63 comprised 1,000 rice, 1,000 other-crop, and 1,000 wetland points, which we used for second-stage
 64 training. Thus, manual interpretation involved checking the 36 polygons across years, rather than
 65 independently labeling 27,000 points. The sampled pixels inherited the labels assigned to their polygons
 66 for the corresponding year and did not require individual interpretation. We have clarified this
 67 distinction and the annual imagery checks in the Methods.

68 We have also compared training configurations using three sample groups to clarify their contributions
 69 to the final accuracy. These groups comprised automatically generated pseudo-labels for rice, other
 70 crops, and wetlands, hard samples extracted from manually interpreted polygons in low-confidence
 71 areas, and background samples automatically extracted from the initial land-cover map to represent
 72 other land covers. We have retained the same GSE features, Random Forest settings, background
 73 samples, and reference locations across these comparisons. Models trained with pseudo-labels and
 74 background samples have achieved F1-scores of 0.756–0.760 across three random seeds, compared
 75 with 0.475–0.483 for models trained with hard samples and background samples. Combining both
 76 sample groups has increased the F1-score to 0.909–0.911 (Figure S8). Recall has ranged from 61.2–
 77 61.6 % with pseudo-labels and background samples and from 31.2–31.9 % with hard samples and
 78 background samples, increasing to 83.9–84.2 % when all three sample groups have been combined.
 79 These results show that pseudo-labels provided information that the targeted hard samples alone did not
 80 capture, and their combination improved rice detection.



81
 82 Figure S8. Accuracy of classifiers trained with pseudo-labels, hard samples, and their combination using the same
 83 GSE features and background samples. Bars and numerical labels show mean precision, recall, and F1-score
 84 across three random seeds, and black points indicate individual runs.

85 We agree that the final-map F1-score of 0.904 represents the complete two-stage framework and cannot
 86 be attributed to pseudo-labels alone. At the same time, the manual contribution consisted of targeted
 87 polygon interpretation followed by automatic sample extraction, rather than individual labeling of
 88 thousands of new points each year. Figure S8 shows that the highest accuracy depended on combining
 89 these targeted samples with pseudo-labels. Because the hard samples were selected from difficult areas

90 and the training configurations differ in sample counts, this comparison supports their complementary
91 roles without assigning a fixed share of the accuracy gain to either source. We have revised the
92 manuscript to distinguish this combined contribution from the performance of the pseudo-label stage
93 alone.

94 **Revised manuscript**

95 Page and line numbers identify the revision locations in the original submission manuscript. Green text shows the
96 targeted revisions. Black text provides surrounding context.

97 **Abstract**

98 [Page 2, Line 10] Here, we present the first country-wide, annual rice paddy maps of Madagascar at 10
99 m resolution from 2017 to 2025. The two-stage Random Forest classification used GSE annual features
100 within a framework of phenology-guided weak supervision with uncertainty-guided refinement. The
101 first stage used automatically generated pseudo-labels, and the second stage incorporated samples
102 extracted from manually interpreted polygons in low-confidence areas. The pseudo-labels formed
103 compact and clearly separated clusters in the GSE feature space across all nine years, and only five GSE
104 dimensions consistently contributed to rice discrimination, indicating that pre-trained embeddings
105 encoded phenologically meaningful information for rice paddy.

106 [Page 2, Line 19] Our maps indicated an increase in mapped rice paddy extent from 886,112 ha in 2017
107 to 1,195,766 ha in 2025 (+34.9 %), with most gains occurring along the margins of existing paddies
108 rather than in new frontiers. This study demonstrates that phenology-guided weak supervision with
109 uncertainty-guided refinement provides a scalable approach for rice paddy mapping in data-scarce
110 regions. The annual rice paddy maps are publicly available from Zenodo at
111 <https://doi.org/10.5281/zenodo.20654510> (Kwon et al., 2026).

112 **Section 1. Introduction**

113 [Page 3, Line 28] Although there have been attempts to employ crowdsourced ground references to
114 produce rice paddy maps in Madagascar, spatially and temporally biased samples can substantially
115 constrain mapping accuracy in countries with heterogeneous landscapes (De Vos et al., 2023). In this
116 context, pseudo-labeling methods generate training samples from physically interpretable satellite
117 signals and provide weak supervision for mapping in data-scarce regions (You et al., 2023; Yang et al.,
118 2021). Nevertheless, the reliability of pseudo-labels has not been rigorously evaluated at the country
119 scale, particularly in ecologically heterogeneous environments such as Madagascar.

120 [Page 4, Line 10] These embeddings have therefore shown strong performance in general land cover
121 classification (Mantey et al., 2025; Li and Yu, 2025). Yet, their potential for crop type mapping,
122 particularly for rice paddy, has not been systematically evaluated. Integrating phenology-guided weak
123 supervision with GSE-based classification and uncertainty-guided refinement could provide a practical
124 rice mapping framework wherever both data sources are available.

125 [Page 4, Line 17] First, we generated pseudo-labels from optical satellite-derived phenological
126 indicators, using the LSWI–NDVI flooding signal to identify rice paddy without relying on extensive
127 ground reference data. Second, we classified GSE annual embeddings using these pseudo-labels and
128 used model confidence to guide the interpretation of difficult areas. We then refined the classifier with
129 samples extracted from the interpreted polygons, where fragmented fields, mixed pixels, or spectrally
130 similar land cover types can reduce classification reliability. Third, we applied harmonic NDVI fitting
131 to characterize cropping intensity over the mapped rice paddy areas.

132 Section 2.2. Overall Framework

133 [Page 5, Line 16] Our rice paddy mapping framework uses phenology-guided weak supervision with
134 uncertainty-guided refinement. It includes land cover classification, rice phenology analysis, rule-based
135 pseudo-label generation, two-stage Random Forest classification, and cropping intensity estimation
136 through harmonic NDVI fitting (Fig. 2). The rationale for separating pseudo-label generation (HLS, 30
137 m) from classification (GSE, 10 m) is that HLS provides the dense temporal observations needed to
138 detect rice-specific phenological signals, while GSE provides higher spatial resolution and gap-free
139 annual features for wall-to-wall mapping.

140 [Page 6, Line 4] To complement the clear rice signals captured by the pseudo-labels, the second
141 classification stage used model confidence to guide polygon interpretation in difficult areas. We
142 extracted hard samples from these polygons to refine the predictions.

143 Section 2.4.4. Rice Paddy Detection

144 [Page 13, Line 27] After the first-stage classification, we used the resulting confidence maps to identify
145 areas where phenology-based pseudo-labels were less sufficient, including fragmented field boundaries,
146 wetland margins, and spectrally mixed cropland. Visual inspection of the initial maps revealed errors
147 concentrated in low-confidence regions, including confusion between rice and wetlands and
148 misclassification of non-cropland areas. To address these errors, the authors, including a food system
149 expert from Madagascar, interpreted 36 polygons representing difficult rice, other-crop, and wetland
150 cases in low-confidence areas. For each year from 2017 to 2025, we reviewed Sentinel-2 RGB imagery
151 and Google Earth imagery, where available, to check for crop rotation, fallow periods, and land-use
152 changes before assigning the polygon labels. We generated 3,000 hard samples per year through
153 stratified random sampling within these polygons at the 10 m resolution of GSE, comprising 1,000
154 samples for each class. The sampled pixels inherited the labels assigned to their polygons for the
155 corresponding year and did not require individual interpretation. We combined these samples with the
156 pseudo-labels and other-land-cover samples automatically extracted through stratified random sampling
157 from the initial land cover map. A Random Forest classifier with 300 trees was then retrained to
158 distinguish rice, other crops, wetlands, and other land covers. We applied the final classifier across the
159 country without reapplying the initial candidate mask, allowing it to identify rice outside the candidate
160 areas.

161 Section 2.5. Accuracy Assessment

162 [Page 15, Line 13] To evaluate classification performance, we compared the predicted maps with the
163 reference samples described in Section 2.3.4. To investigate the contributions of the training samples,
164 we have compared models trained with pseudo-labels, hard samples, and their combination using the
165 same GSE features, background samples, Random Forest settings, and reference locations. The
166 background samples represented other land covers and were automatically extracted from the initial
167 land cover map. Overall accuracy (OA) was calculated as the proportion of correctly classified samples
168 among all validation samples.

169 Section 3.1. Accuracy assessment of rice paddy classification

170 [Page 17, Line 13] In contrast, the 2nd-stage classification missed 168 rice points and produced only 8
171 false rice predictions. The comparison of training configurations has further clarified the contribution
172 of each sample source, yielding F1-scores of 0.756–0.760 for pseudo-labels with background samples,
173 0.475–0.483 for hard samples with background samples, and 0.909–0.911 for their combination (Fig.
174 S8). The highest accuracy has resulted from combining the two sample sources, supporting their
175 complementary contributions to rice detection. Across Köppen–Geiger climate zones, F1-scores ranged
176 from 0.84 in the rainforest zone to 0.96 in the steppe zone, with precision ranging from 0.98 to 1.00
177 (Fig. 6b).

178 Section 4.2. Pre-trained embeddings captured rice phenology in a compact feature space

179 [Page 28, Line 17] Because GSE integrates information from multiple sensors and summarizes broader
180 spectral and temporal patterns, it provided a compact input that helped reduce reliance on cloud-gap
181 filling and manual feature engineering. Our targeted refinement helped the classifier use the information
182 in GSE embeddings to distinguish difficult rice and non-rice cases. Errors along field edges and wetland
183 margins in the first-stage map indicate that a compact feature space alone does not ensure accurate
184 classification where land covers share similar characteristics (Fig. S5). Adding samples from these
185 difficult areas improved classification while preserving the broader information supplied by pseudo-
186 labels (Fig. S8). This suggests that the usefulness of pre-trained embeddings depends on training
187 samples that represent both clear phenological patterns and ambiguous land-cover conditions.
188 Combining these sample sources therefore helped translate the information captured by GSE into more
189 reliable rice predictions without constructing additional crop-specific features.

190 Section 5. Conclusions

191 [Page 30, Line 14] This study demonstrates a scalable approach for annual rice paddy mapping in data-
192 scarce regions through phenology-guided weak supervision with uncertainty-guided refinement using
193 pre-trained satellite embeddings. Rather than relying on dense field reference data, the framework uses
194 the seasonal flooding-to-greenup sequence of transplanted rice to generate training samples and uses
195 GSE to represent rice-related phenology in a compact feature space.

196

197 **Major Comment 2.**

198 *The manuscript states that the independent validation samples were manually interpreted using high-*
199 *resolution Sentinel-2 RGB imagery and HLS-derived NDVI/LSWI time series, supported by local*
200 *expertise. The second-stage hard samples were also manually interpreted using high-resolution imagery*
201 *and NDVI/LSWI time-series information, and they were specifically selected from low-confidence or*
202 *error-prone areas identified by the first-stage model. Because the training refinement and validation*
203 *procedures rely on highly similar interpretation sources, strict separation between training and*
204 *evaluation is essential. At present, the manuscript does not provide enough information to determine*
205 *whether the hard samples and validation points are spatially and temporally independent, whether the*
206 *same interpreters or reference information were used, or whether validation results influenced hard-*
207 *sample selection. Consequently, it is difficult to assess whether the reported OA of 91.2%, precision of*
208 *99.0%, recall of 83.2%, and F1-score of 0.904 represent a fully independent estimate of generalization*
209 *performance.*

210 *Response: Thank you very much for highlighting the importance of separating second-stage training*
211 *from validation. We have clarified the roles of the interpreters and how the two processes were*
212 *conducted. The authors responsible for training refinement, including a food system expert from*
213 *Madagascar, jointly interpreted the hard-sample polygons. Hervet Randriamady, a co-author and local*
214 *expert from Madagascar, interpreted only the rice and other-crop reference samples and did not*
215 *participate in hard-sample interpretation. Figure R1 shows examples of the HLS-derived NDVI and*
216 *LSWI time series that he considered alongside the imagery when assigning these reference labels. We*
217 *prepared the labels after rice paddy map production without consulting the final rice map. The reference*
218 *samples were not used for model training, and validation results did not inform hard-sample selection.*

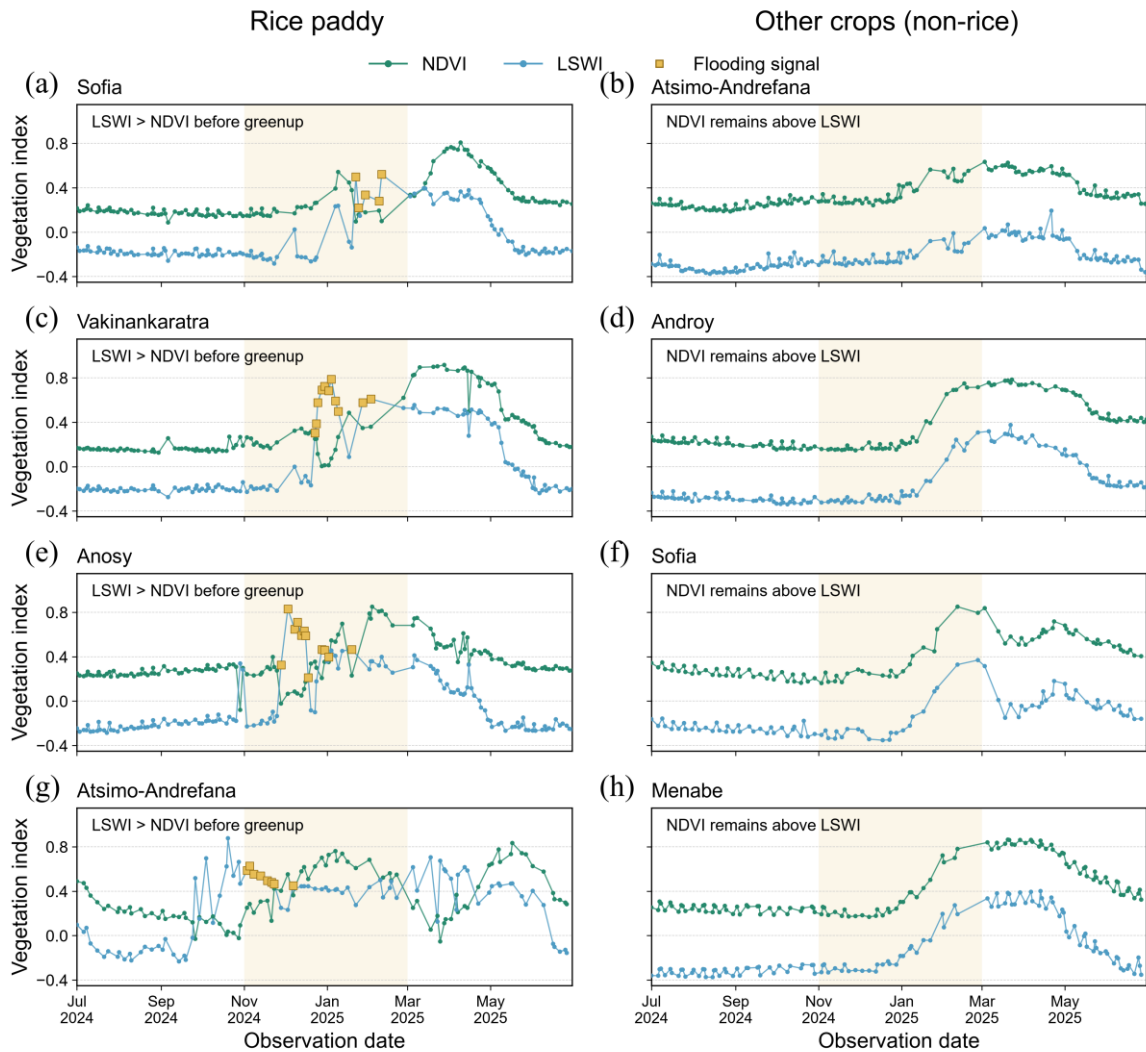
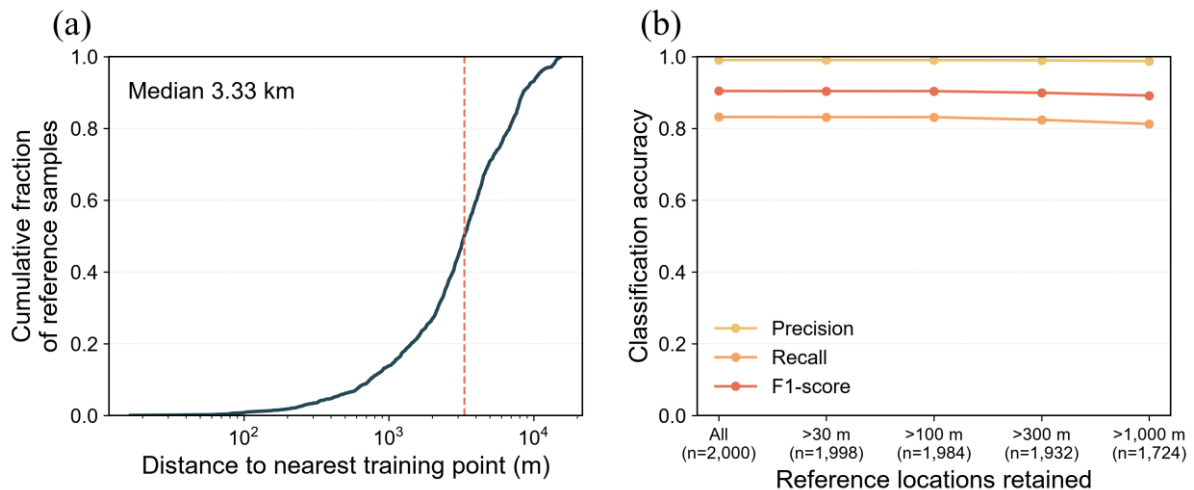


Figure R1. HLS-derived NDVI and LSWI time series at four rice-paddy reference locations (a, c, e, g) and four other-crop reference locations (b, d, f, h). Green and blue indicate NDVI and LSWI, respectively. Shaded areas indicate the typical November–February transplanting period, and yellow squares mark observations with LSWI above NDVI before crop growth.

We have further evaluated the spatial separation between training and reference samples. Among the 2,000 reference locations, including 1,000 rice and 1,000 other-crop locations, none falls within a hard-sample polygon or shares a mapped pixel with a training point. The minimum distance to the nearest training point is 16.6 m, and the median distance is 3.33 km (Figure S9a). We have also repeated the accuracy assessment after excluding reference locations within 30, 100, 300, and 1,000 m of training points. The corresponding F1-scores are 0.904, 0.904, 0.899, and 0.892 (Figure S9b). These small differences show that the reported accuracy does not depend strongly on reference locations close to training points.



233

234 Figure S9. Spatial separation of training and reference samples and sensitivity of final-map accuracy to nearby
 235 training points. (a) Cumulative distribution of the distance to the nearest training point. (b) Precision, recall, and
 236 F1-score after excluding reference locations within each distance threshold. Sample counts indicate the retained
 237 reference locations.

238 Both interpretation processes used satellite imagery and HLS-derived NDVI and LSWI time series.
 239 Although the interpreters and labeling procedures were separate, these shared information sources may
 240 lead to similar interpretation errors. We have therefore clarified that the reported accuracy describes
 241 performance at the separately interpreted reference locations, with the spatial separation supported by
 242 Figure S9. Additional reference data, covering other regions, and supported by independent observation
 243 sources would strengthen the assessment of generalization and interpretation uncertainty.

244 Revised manuscript

245 Page and line numbers identify the revision locations in the original submission manuscript. Green text shows the
 246 targeted revisions. Black text provides surrounding context.

247 Section 2.3.4. Validation Datasets

248 [Page 9, Line 22] We evaluated the land cover classification and rice paddy detection results using
 249 independently generated reference samples. For rice paddy detection, a local expert from Madagascar
 250 interpreted the rice-paddy and other-crop reference samples using high-resolution Google Earth RGB
 251 imagery, HLS-derived NDVI and LSWI time series, and local knowledge of rice cultivation. Clearly
 252 defined field boundaries and repeated field patterns supported identification of agricultural land, and
 253 we examined homogeneous interiors to avoid mixed boundaries. Rice-paddy interpretation combined
 254 dark-green, wet-looking fields with LSWI temporarily exceeding NDVI during transplanting and a
 255 subsequent increase in NDVI. Other-crop interpretation combined drier cultivated fields with NDVI
 256 remaining clearly above LSWI through the annual observation period. We considered the imagery and
 257 the seasonal relationship between both indices together when assigning labels.

[Page 9, Line 22] Additionally, the reference interpretation and training refinement were separate processes. The reference labels were prepared after map production. The reference interpreter did not participate in hard-sample interpretation or consult the final rice map. We did not use reference samples for training or validation results to select hard samples. After removing duplicate coordinates, the reference dataset contained 1,000 rice-paddy and 1,000 other-crop locations for the assessment. We have made the reference coordinates and class labels publicly available at <https://doi.org/10.5281/zenodo.23292193>.

Section 2.5. Accuracy Assessment

[Page 15, Line 2] To evaluate classification performance, we compared the predicted maps with the reference samples described in Section 2.3.4. We evaluated the final rice paddy map against the separately interpreted rice and other-crop reference samples using overall accuracy, precision, recall, and F1-score. To assess whether the distance between training and reference samples affected the accuracy estimates, we have checked for shared locations and repeated the evaluation after excluding reference locations within 30, 100, 300, and 1,000 m of any training point. Overall accuracy (OA) was calculated as the proportion of correctly classified samples among all validation samples.

Section 3.1. Accuracy assessment of rice paddy classification

[Page 17, Line 2] In contrast, the 2nd-stage classification missed 168 rice points and produced only 8 false rice predictions. The final-map accuracy has remained similar after excluding reference locations near training points, with F1-scores of 0.892–0.904 across exclusion distances of 30–1,000 m (Fig. S9). Across Köppen–Geiger climate zones, F1-scores ranged from 0.84 in the rainforest zone to 0.96 in the steppe zone, with precision ranging from 0.98 to 1.00 (Fig. 6b).

Section 4.4. Implications for future applications

[Page 30, Line 6] In addition, shared interpretation sources remain a limitation even when training and reference labeling use separate interpreters and procedures. Additional reference data from other regions, supported by independent observation sources, would help assess how well the framework generalizes and reduce uncertainty in reference interpretation.

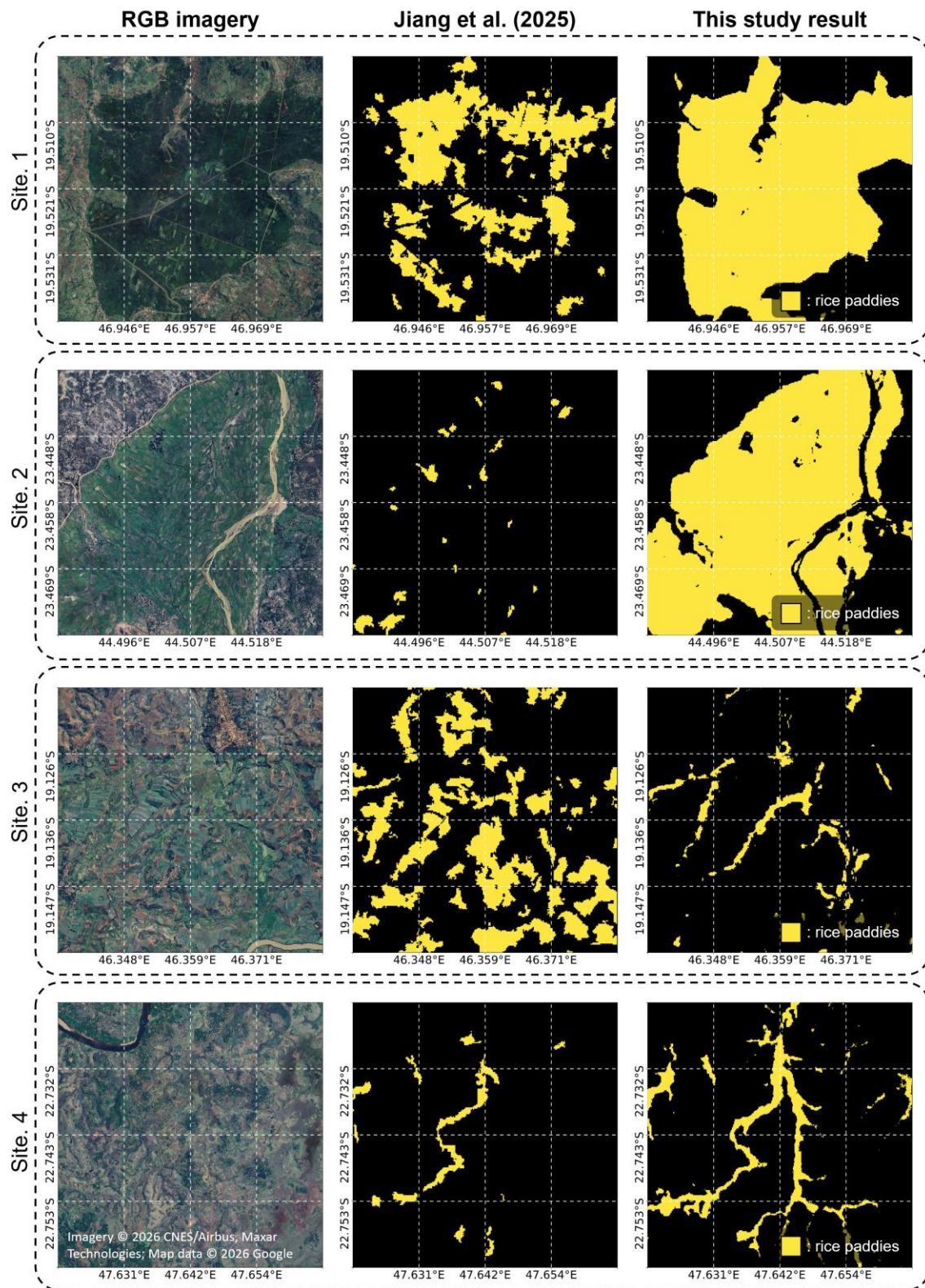
Major Comment 3.

The manuscript compares the proposed 10 m annual rice maps with the 20 m Africa rice distribution map for 2023 and repeatedly refers to this dataset as a “SAR-based benchmark.” However, according to the manuscript’s own description, the benchmark product was generated using both Sentinel-1 SAR and Sentinel-2 optical data, together with object-based segmentation and Random Forest classification. Referring to it simply as “SAR-based” may therefore be misleading. In addition, the proposed product covers 2017–2025 at 10 m, whereas the benchmark is a 2023 product at 20 m. The manuscript does not sufficiently explain how spatial resolution, temporal scope, validation sampling, and target-class definitions were harmonized before calculating the very large performance difference. The subsequent attribution of the benchmark’s poorer performance mainly to SAR viewing geometry, terrain shadow, overlapping orbits, and backscatter confusion is also difficult to isolate from other differences in training data, algorithm design, spatial resolution, and class definition. These factors make the fairness of the comparison and the causal interpretation of the benchmark’s lower accuracy uncertain.

Response: Thank you very much for highlighting the need for a comparable assessment and a clearer interpretation of the differences between the rice maps. We agree that “SAR-based benchmark” does

299 not fully describe the method used by Jiang et al. (2025). Their framework combined Sentinel-2 optical
300 information for object-based segmentation with Sentinel-1 SAR features for Random Forest
301 classification. We have therefore replaced this wording with a description of the rice map from Jiang et
302 al. (2025) and renamed Section 2.3.5 “Existing rice mapping products.”

303 The original accuracy assessment compared our 2023 map with the 2023 map from Jiang et al. (2025)
304 at the same reference locations. In the original manuscript, we extracted predictions at the native
305 resolution of each product, which was 10 m for our map and 20 m for Jiang et al. (2025). We treated
306 rice as the positive class and grouped our other-crop, wetland, and other-land-cover classes as non-rice
307 to match their binary map. The reference labels distinguished rice from other crops, so the accuracy
308 estimates described discrimination between these reference classes. Figure S4 complemented this
309 assessment by illustrating mapped rice at four sites.

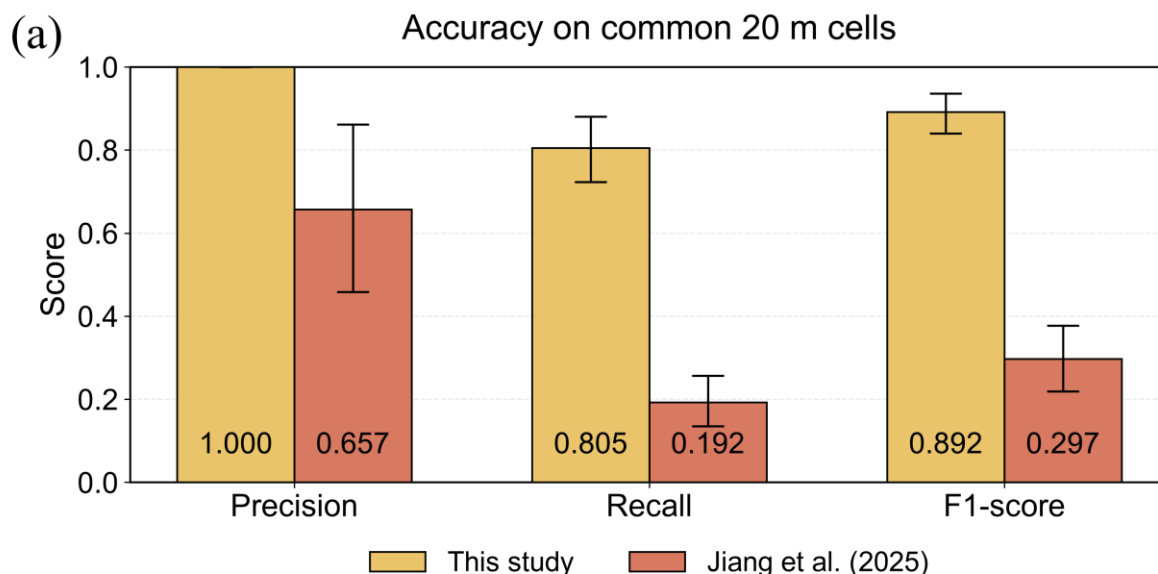
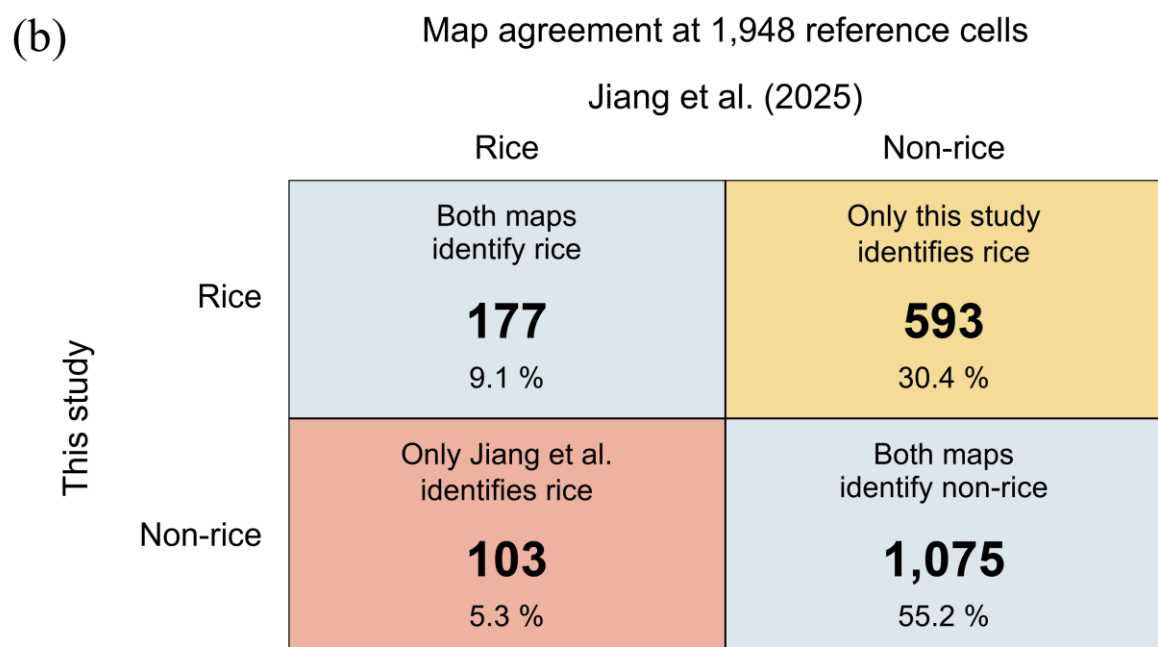


310

311 Figure S4. Visual comparison of the 2023 rice maps from Jiang et al. (2025) and this study at four sites in
 312 Madagascar. The left column shows Google Satellite RGB imagery, and the middle and right columns show
 313 mapped rice in yellow from Jiang et al. (2025) and this study, respectively, at their native resolutions (background
 314 imagery © 2026 CNES/Airbus, Maxar Technologies. Map data © 2026 Google).

315 We also have added a common-grid comparison to assess whether the accuracy difference remains after
316 matching spatial resolution. We have used both 2023 maps on common 20 m cells at the same reference
317 locations. We have preserved cells with at least 95 % valid coverage in both maps and classified a cell
318 as rice when mapped rice covers at least half of its area. After counting each cell once, the comparison
319 includes 1,948 cells with 957 rice and 991 other-crop reference locations. We have retained the original
320 point-based reference labels for this assessment.

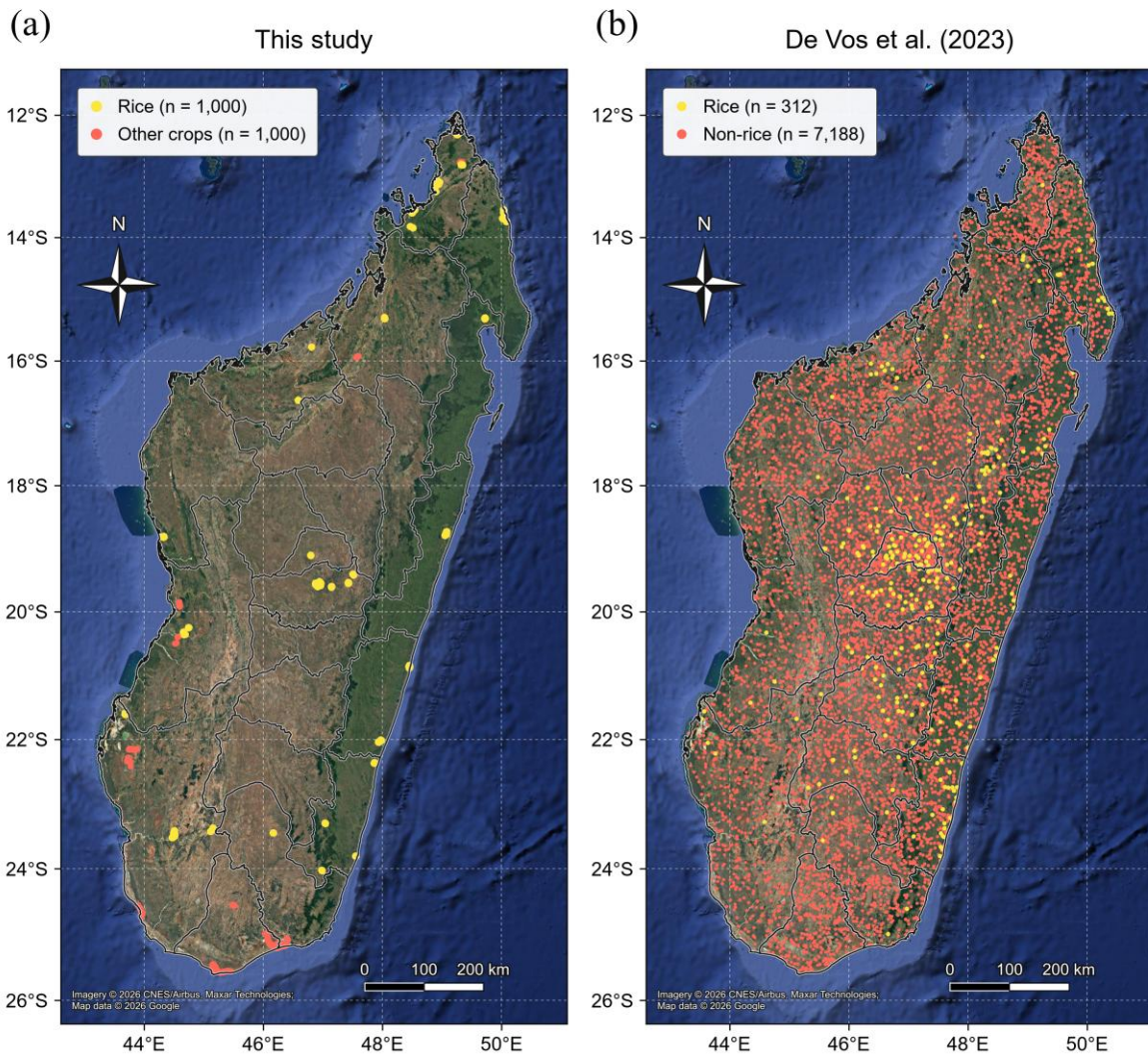
321 Our map has achieved precision of 100.0 %, recall of 80.5 %, and an F1-score of 0.892, compared with
322 65.7 %, 19.2 %, and 0.297, respectively, for Jiang et al. (2025) on the common grid (Figure S11a). The
323 F1-score difference has remained positive when we have resampled the paired predictions in 10 km
324 spatial blocks, with a conditional 95 % interval of 0.502–0.691. These results show that matching spatial
325 resolution has not removed the accuracy difference on our reference dataset. Figure S11b separately
326 shows where the maps agree or only one identifies rice.



327

328 Figure S11. Comparison of the 2023 rice maps on common 20 m cells. (a) Precision, recall, and F1-score against
 329 the rice and other-crop reference labels. Error bars show conditional 95 % intervals from 10 km spatial-block
 330 resampling. (b) Numbers and percentages of the 1,948 cells where both maps agree or only one map identifies
 331 rice.

332 We have also compared the three products using our reference data and the independent test data
 333 provided by De Vos et al. (2023), following the comparison settings described in our response to
 334 Community Comment 1. De Vos et al. (2023) used multi-annual Sentinel-1 and Sentinel-2 observations
 335 from 2017 to 2020 for rice mapping and interpreted their reference labels using very-high-resolution
 336 imagery. We have evaluated all three maps at the same locations within each reference dataset, including
 337 2,000 locations from this study and 7,459 locations from De Vos et al. (2023) (Figure S1).



338

339 Figure S1. Spatial distribution of reference samples in Madagascar from (a) this study and (b) De Vos et al. (2023).
 340 Yellow points indicate rice, and red points indicate other crops in (a) and non-rice in (b) (background imagery:
 341 Imagery © 2026 CNES/Airbus, Maxar Technologies; Map data © 2026 Google).

342 On our reference dataset, our map has achieved an F1-score of 0.904, compared with 0.805 for De Vos
 343 et al. (2023) and 0.298 for Jiang et al. (2025) (Figure S6a). On the reference data from De Vos et al.
 344 (2023), our map has achieved precision of 60.9 %, recall of 35.9 %, and an F1-score of 0.452, compared
 345 with 58.9 %, 40.4 %, and 0.479, respectively, for De Vos et al. (2023), and 36.7 %, 12.8 %, and 0.190
 346 for Jiang et al. (2025) (Figure S6b). Our map has therefore shown higher precision, recall, and F1-score
 347 than Jiang et al. (2025) on both reference datasets, while rice omissions remain evident on the reference
 348 data from De Vos et al. (2023). These comparisons provide additional context for the common-grid
 349 assessment in Figure S11, although they include products from different mapping periods.

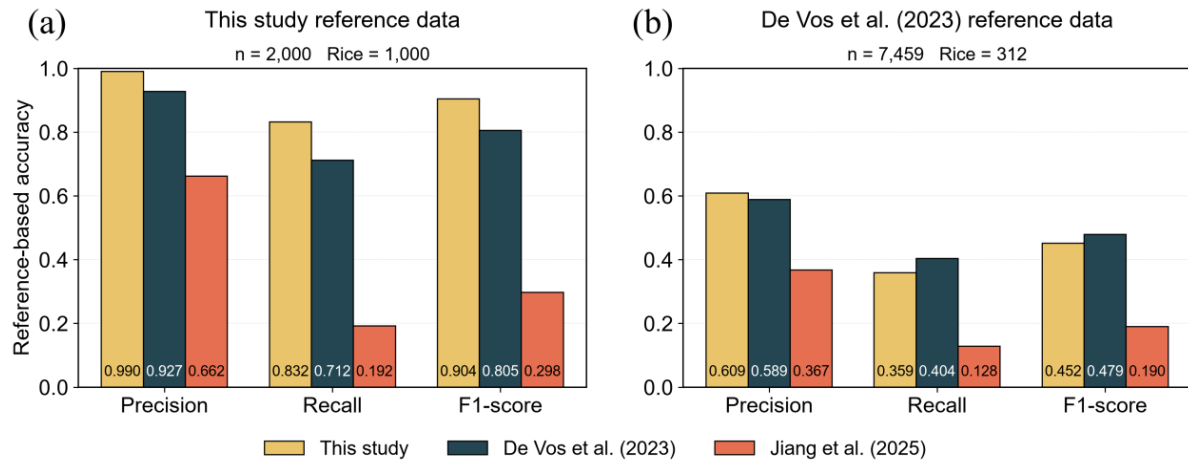


Figure S6. Precision, recall, and F1-score of rice maps from this study, De Vos et al. (2023), and Jiang et al. (2025), evaluated using reference data from (a) this study and (b) De Vos et al. (2023).

We have also evaluated whether terrain and Sentinel-1 observation conditions support the original explanation of the accuracy difference. The rice omission rate for Jiang et al. (2025) has reached 80.9 % at reference locations with slopes below 2° and 83.0 % at locations with slopes of at least 5°. This small difference does not support steep terrain as the main explanation for the rice omissions. Some orbit conditions have been associated with more false rice detections, but the corresponding reference locations occur in different geographic areas. These comparisons therefore do not separate orbit effects from regional differences or establish terrain shadow as a cause.

A further comparison has shown that SAR features preserve useful information for rice discrimination. Using the same reference dataset and Random Forest settings with spatially separated training and test locations, we have obtained F1-scores of 0.926 for Sentinel-1 features, 0.956 for Sentinel-2 features, and 0.966 for their combination. This comparison assesses the information in the features and does not reproduce the trained model of Jiang et al. (2025). It therefore does not establish why their map differs from ours or demonstrate a general advantage of optical over SAR data.

We have consequently revised the Discussion to focus on the observed differences between the maps and removed the attribution of the lower accuracy mainly to SAR viewing geometry, terrain shadow, overlapping orbits, and backscatter confusion. Our phenological rules identified transplanted rice through flooding followed by vegetation growth. Although we grouped both maps into rice and non-rice, the two methods may identify different types of rice cultivation. The products also differ in the seasons covered by satellite observations, the training samples, and the classification methods. Our additional analyses have provided a more consistent comparison, but they cannot determine how much of the observed difference results from the choice of satellite sensor.

374 Revised manuscript

375 Page and line numbers identify the revision locations in the original submission manuscript. Green text shows the
376 targeted revisions. Black text provides surrounding context.

377 Abstract

378 [Page 2, Line 15] The pseudo-labels formed compact and clearly separated clusters in the GSE feature
379 space across all nine years, and only five GSE dimensions consistently contributed to rice
380 discrimination, indicating that pre-trained embeddings encoded phenologically meaningful information
381 for rice paddy. The final map achieved an overall accuracy (OA) of 91.2 %, precision of 99.0 %, recall
382 of 83.2 %, and an F1-score of 0.904 on the rice and other-crop reference samples. Our maps indicated
383 an increase in mapped rice paddy extent from 886,112 ha in 2017 to 1,195,766 ha in 2025 (+34.9 %),
384 with most gains occurring along the margins of existing paddies rather than in new frontiers.

385 Section 2.3.5. Existing rice mapping products

386 [Page 10, Line 11] We compared our 2023 rice paddy map with the 2023 Africa-wide rice map produced
387 by Jiang et al. (2025). This product covers 34 rice-producing African countries, including Madagascar.
388 It combined Sentinel-2 optical information for object-based segmentation with Sentinel-1 SAR features
389 for Random Forest classification. We first evaluated both maps at their native resolutions, grouping
390 mapped classes as rice and non-rice. To account for their different spatial resolutions, we have also
391 compared both maps on common 20 m cells, retaining cells with at least 95 % valid coverage in each
392 product and identifying rice where it covered at least half of the cell.

393 Section 2.5. Accuracy Assessment

394 [Page 15, Line 13] To evaluate classification performance, we compared the predicted maps with the
395 reference samples described in Section 2.3.4. We have also accounted for spatial dependence in the
396 comparison with Jiang et al. (2025). For this comparison, we have evaluated both maps on a common
397 grid at 1,948 unique cells using the original point-based reference labels and assessed uncertainty in the
398 F1-score difference by resampling the paired predictions in 10 km spatial blocks. Overall accuracy (OA)
399 was calculated as the proportion of correctly classified samples among all validation samples.

400 Section 3.1. Accuracy assessment of rice paddy classification

401 [Page 17, Line 13] In contrast, the 2nd-stage classification missed 168 rice points and produced only 8
402 false rice predictions. The accuracy difference between the two 2023 maps has also remained after
403 matching spatial resolution, with F1-scores of 0.892 for our map and 0.297 for Jiang et al. (2025) on
404 common 20 m cells (Fig. S11). The difference has remained positive under spatial-block resampling,
405 with a conditional 95 % interval of 0.502–0.691. Across Köppen–Geiger climate zones, F1-scores
406 ranged from 0.84 in the rainforest zone to 0.96 in the steppe zone, with precision ranging from 0.98 to
407 1.00 (Fig. 6b).

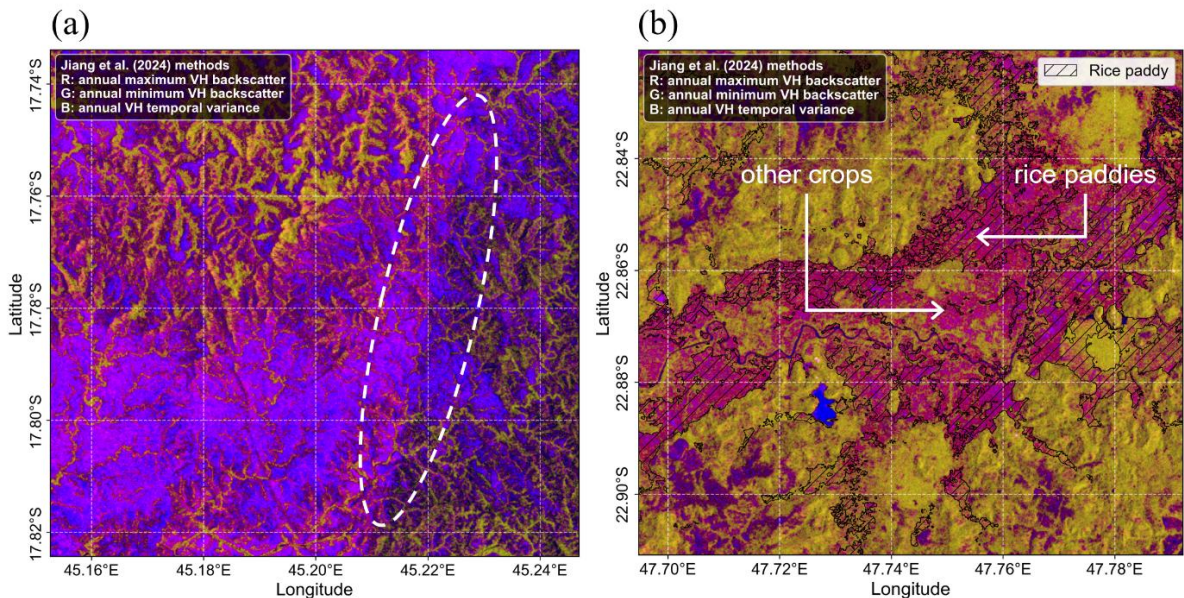
408 **Section 4.1. Phenology-guided pseudo-labels supported annual rice mapping**

409 [Page 27, Line 3] Phenology-guided pseudo-labels supported annual rice mapping by complementing
410 targeted manual refinement. This combination is particularly useful in Madagascar, where limited
411 reference data and diverse rice systems make representative training challenging. De Vos et al. (2023)
412 showed that crowdsourced observations mainly captured visually distinctive, intensive rice systems and
413 did not adequately represent conditions across the country. Our framework generated high-confidence
414 pseudo-labels from satellite-observed phenology and used model confidence to guide interpretation in
415 difficult areas. Pseudo-labels with background samples achieved higher F1-scores than hard samples
416 with the same background data, while their combination produced the highest accuracy (Fig. S8). These
417 results support their complementary use for detailed rice detection without a new nationwide field
418 campaign (Fig. 5 and Fig. 6). The pseudo-labeling rules were effective because they described the
419 seasonal sequence of transplanted rice rather than a single spectral condition.

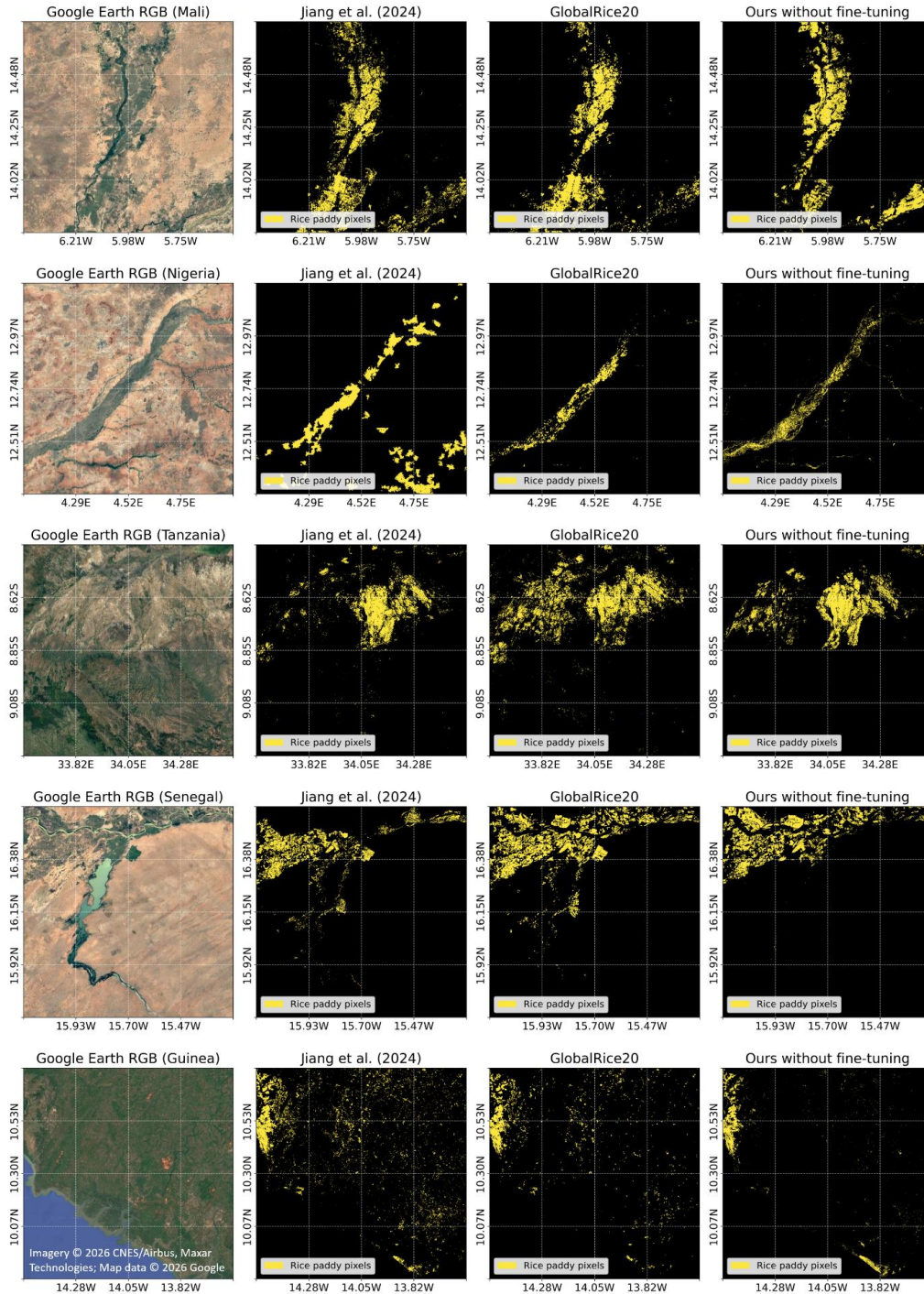
420 [Page 27, Line 25] The common-grid comparison showed that the framework’s accuracy advantage
421 extended beyond its finer spatial resolution. On common 20 m cells, our 2023 map retained an F1-score
422 of 0.892, compared with 0.297 for Jiang et al. (2025), indicating that resolution alone did not explain
423 the difference (Fig. S11). Recent studies have similarly demonstrated the value of combining
424 phenological knowledge with complementary satellite information for automated sample generation
425 and rice mapping (Yang et al., 2024; Zhao et al., 2024). Our results extend this approach by combining
426 phenology-guided pseudo-labels, annual embeddings, and targeted refinement for repeated national
427 mapping. The training comparisons and local examples support the practical value of this combination
428 where reference data are limited (Fig. S8 and Fig. S4).

429 **Revised supplementary captions**

430 Figures S8, S9, and S1 are shown above and are not repeated here.



432 Figure S13. Sentinel-1 VH backscatter patterns at two sites in Madagascar. Background images show annual
433 maximum, annual minimum, and annual temporal variance of VH backscatter in the red, green, and blue channels,
434 respectively. Black hatched polygons indicate pixels classified as rice paddies in this study.



435

436 Figure S16. Exploratory spatial transferability assessment of the Madagascar-trained rice paddy mapping
 437 framework in African rice-producing regions. Each row shows Google Earth RGB imagery, the rice map from
 438 Jiang et al. (2025), GlobalRice20 (Zhang et al., 2026), and our result without country-specific fine-tuning. Yellow
 439 indicates mapped rice paddies (background imagery © 2026 CNES/Airbus, Maxar Technologies. Map data ©
 440 2026 Google).

441 **Reviewer 2**

442 *The major contribution of this study is the development of annual rice paddy maps for Madagascar, an*
443 *important rice-producing country in Africa where high-resolution rice datasets remain relatively limited.*
444 *In my view, any effort to improve spatially explicit rice information in Africa deserves attention,*
445 *particularly when the resulting products are made publicly available. However, I still have several*
446 *concerns regarding the method used to estimate cropping intensity and several other methodological*
447 *aspects. My specific comments are as follows.*

448 *Response: Thank you for recognizing the value of the annual Madagascar rice maps and for the specific*
449 *suggestions on cropping intensity and methodology. We have clarified what the NDVI peak counts*
450 *represent and how they differ from the number of rice cultivation cycles. We have also provided more*
451 *detail on how we interpreted the reference samples and assessed how many rice reference locations fall*
452 *within the initial candidate mask. In addition, we have revised the description of manual refinement,*
453 *renamed the comparison product, and removed general textbook explanations from the Methods. Below,*
454 *we explain how we have addressed each comment and revised the manuscript accordingly.*

455

456 **Major Comment 1.**

457 *My major concern is the definition and estimation of cropping intensity. If I understand the method*
458 *correctly, the authors first identify rice pixels and then estimate cropping intensity by fitting the HLS*
459 *NDVI time series and counting vegetation peaks. Pixels with one NDVI peak are classified as single-*
460 *cropped rice, whereas those with two peaks are classified as double-cropped rice. However, what is*
461 *actually detected here is the number of vegetation growth cycles within pixels classified as rice, rather*
462 *than the number of rice cultivation cycles themselves. These two concepts are not necessarily equivalent.*
463 *Even if rice-wheat rotation or rice-other crop rotations are uncommon in Madagascar, I do not think*
464 *that rice cropping intensity can be defined directly from the number of NDVI peaks. Each identified rice*
465 *cycle should be supported by rice-specific phenological evidence rather than vegetation peaks alone. I*
466 *therefore suggest that the authors reconsider and improve the current method used to estimate cropping*
467 *intensity.*

468 *Response: Thank you very much for highlighting the distinction between vegetation growth cycles and*
469 *rice cultivation cycles. We agree that NDVI peak counts should not be interpreted directly as the number*
470 *of rice cultivation cycles. Our original analysis aimed to characterize cropping intensity within rice*
471 *paddies, including rice followed by another crop, rather than rice–rice cultivation alone. We*
472 *acknowledge that the wording “rice cropping intensity” did not clearly convey this distinction and have*
473 *clarified the intended meaning.*

474 *Cropping intensity describes the number of crop cultivation cycles within a year, without requiring the*
475 *same crop in each cycle. Previous remote sensing studies have used this definition to characterize annual*
476 *cropping patterns (L. Liu et al., 2020; Pan et al., 2021). Sequential double-cropping can therefore*
477 *include both repeated cultivation of the same crop and cultivation of different crops on the same field*
478 *(Waha et al., 2020). Our previous study provides a related example. In Brazil, we mapped second-season*
479 *maize grown after soybean harvest, so double-cropping referred to the combined soybean–maize*
480 *sequence rather than two maize crops (Kwon et al., 2026). This broader meaning also underlies our*
481 *analysis of cropping intensity within Madagascar rice paddies.*

Our workflow used rice phenology and NDVI seasonality for different purposes. The phenological rules used for pseudo-label generation targeted the flooding-to-greenup sequence associated with rice cultivation. We then analyzed seasonal NDVI dynamics within the mapped rice paddies to estimate cropping intensity. A second flooding signal would be relevant to identifying a second rice cycle, but would not be required for a subsequent non-rice crop. We have preserved this distinction and have added examples showing one or two distinct growth periods together with a pre-growth LSWI greater than NDVI signal (Figure 4). These examples illustrate how rice-related phenology and seasonal growth provide complementary information.

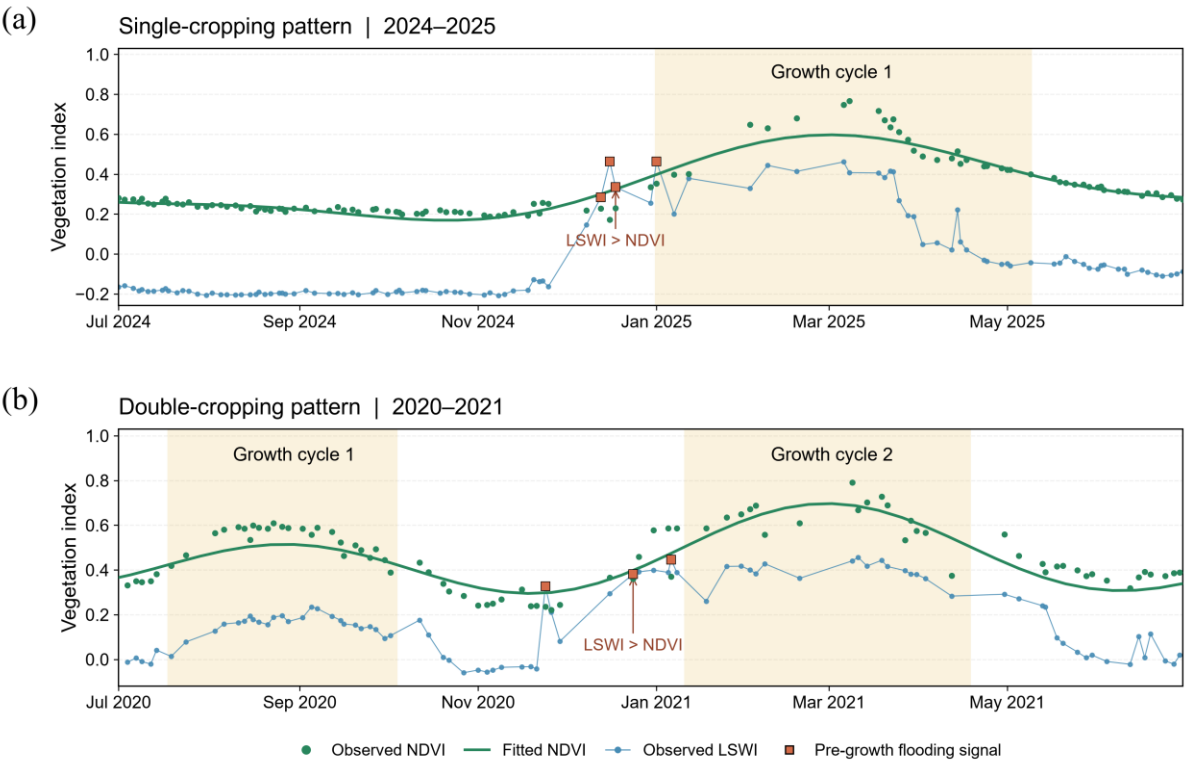


Figure 4. Illustrative HLS time series showing (a) single- and (b) double-cropping patterns within mapped rice paddies. Green points and lines show observed and fitted NDVI, respectively. Shading areas indicate growth periods around the fitted peaks.

Nevertheless, we agree that two NDVI peaks alone provide insufficient evidence of separate cultivation cycles. We have therefore improved the estimation method by considering the seasonal development surrounding each peak. Previous studies have identified crop cycles using NDVI peak values, the time interval between successive peaks, growth-period duration, and NDVI differences between each peak and the low-NDVI periods before and after it (L. Liu et al., 2020; C. Liu et al., 2020; Guo et al., 2021; Ren et al., 2024). Following these principles, we have retained second-order harmonic fitting and added explicit conditions for peak height, separation, prominence, and width. We have also supplemented the harmonic curve with local Savitzky–Golay curves to recover separate growth periods that a smooth annual curve can merge. These additional cycles require both rising and declining NDVI in the fitted curves and supporting observations around the peaks and adjacent low-NDVI periods. The table below summarizes the changes.

Table R1. Comparison of the original and revised methods for estimating cropping intensity within mapped rice paddies.

Aspect	Original method	Revised method
Interpretation	Estimate rice cultivation cycles.	Estimate total cultivation cycles within rice paddies, including rice followed by another crop.
Curve fitting	Fit second-order harmonic regression curves.	Retain harmonic regression and add local Savitzky–Golay curves to distinguish separate growth periods.
Peak screening	Count local NDVI peaks of at least 0.25.	Require harmonic peaks of at least 0.35, at least 80 days between successive peaks, and sufficient prominence and width.
Evidence for a second cycle	Identify a second cycle from fitted peak counts.	Check local growth periods against observed NDVI rise and decline to support a second cycle.
Year boundaries	Count peaks within the analysis year.	Include 60 days before and after the analysis year and assign cycles by peak date.

We have assessed how well the revised rule classifies human-interpreted cropping patterns at 343 locations, including 186 single-cropping and 157 double-cropping locations. We have used these locations to select the rule and calculate classification agreement. Agreement for single cropping has increased from 73.1 % to 92.5 %, and agreement for double cropping has increased from 82.8 % to 86.6 % (panel (a) of the figure below). The revised rule has correctly classified 308 locations, compared with 266 for the original map. Overall agreement has consequently increased from 77.6 % to 89.8 %, and macro-F1 has increased from 0.775 to 0.897 (Figure R2b). These results show improved classification of both cropping patterns on the available labeled locations.

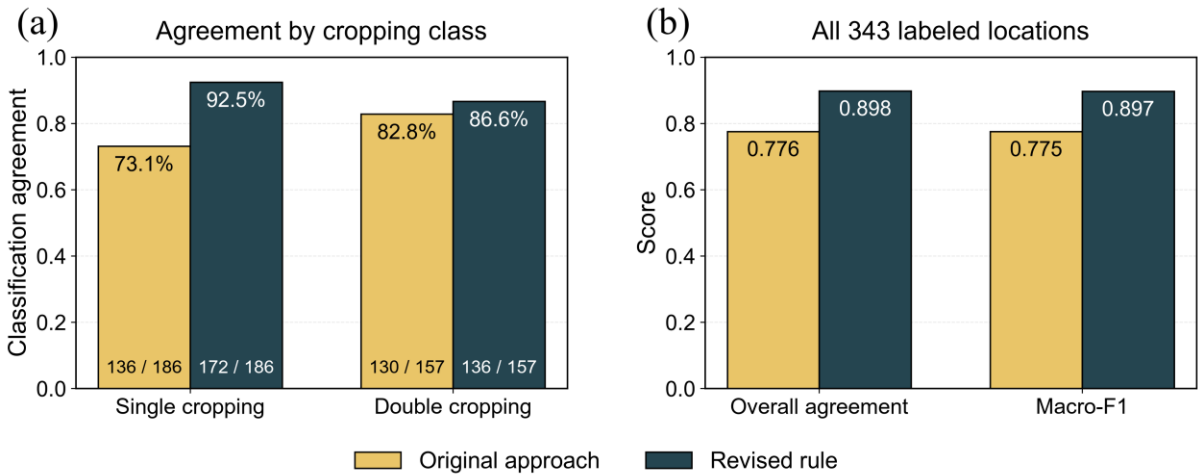
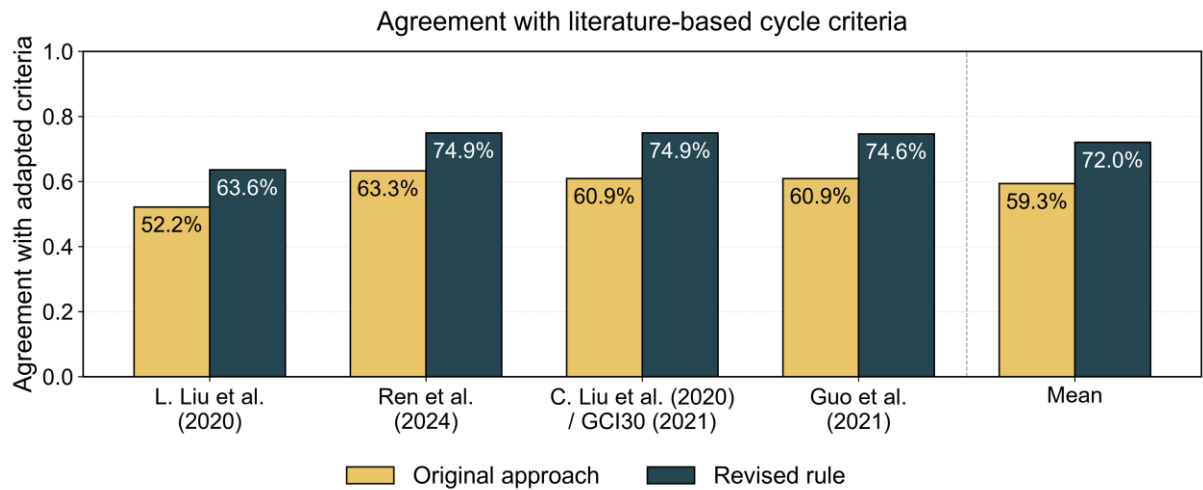


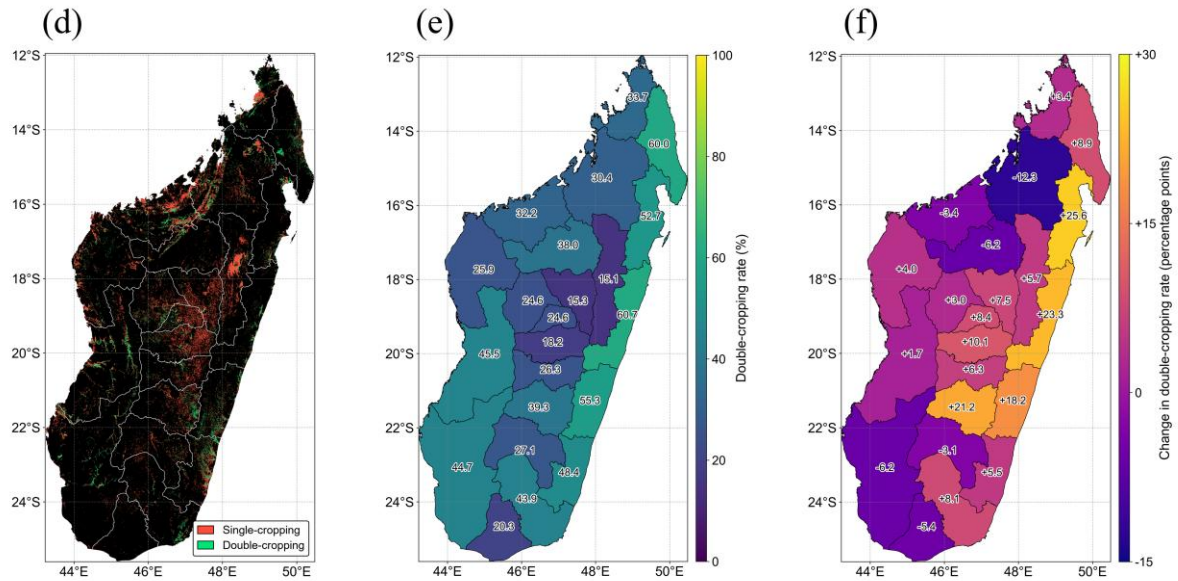
Figure R2. Classification agreement of the original map (ours, before revision) and revised rule (ours, after revision) with human-interpreted cropping labels. (a) Agreement for single and double cropping, with correctly classified locations shown as counts. (b) Overall agreement and macro-F1 across all 343 locations.

521 We have also compared both approaches with four sets of cropping-cycle criteria adapted from the
 522 literature. We have applied these criteria to the same reconstructed NDVI time series at all 343 locations.
 523 Mean agreement has increased from 59.3 % for the original map to 72.0 % for the revised rule, with
 524 improvement for all four sets of criteria (Figure R3). The comparison includes the criteria from L. Liu
 525 et al. (2020), Ren et al. (2024), C. Liu et al. (2020) and GCI30 (Zhang et al., 2021), and Guo et al.
 526 (2021). These results support greater consistency with published approaches to identifying crop growth
 527 cycles.



528
 529 Figure R3. Agreement of the original map and revised rule with four adapted literature criteria at all 343 locations.
 530 Agreement describes consistency with adapted criteria rather than independent accuracy.

531 We have applied the revised rule to the 2017 and 2025 rice-paddy maps and recalculated the cropping-
 532 intensity results corresponding to Figure 11d–f. The estimated national double-cropping rate has
 533 increased from 31.2 % in 2017 to 32.5 % in 2025. Regional rates in 2025 have ranged from 15.1 % to
 534 60.7 %, with higher rates in several eastern regions than in the central highlands. Regional changes
 535 between the two years have ranged from –12.3 to +25.6 %.



536

537 Figure 11d–f. Cropping-intensity patterns within Madagascar rice paddies. (d) Potential single- and double-
 538 cropping systems in 2025, (e) regional double-cropping rates in 2025, and (f) changes between 2017 and 2025.
 539 Double-cropping includes rice followed by another crop and does not necessarily indicate two rice cultivation
 540 cycles. Rates represent double-cropped area as a percentage of the area classified as single-cropping or double-
 541 cropping.

542 Together, the revised rules, improved agreement with human-interpreted labels, and comparison with
 543 published cycle criteria provide stronger support for estimating cropping intensity within rice paddies.
 544 We have preserved the interpretation of potential single- and double-cropping, which includes
 545 sequential cultivation of rice and other crops, rather than confirmed numbers of rice cultivation cycles.
 546 This revision addresses both the meaning of the indicator and the evidence used to distinguish separate
 547 growth periods.

548 Revised manuscript

549 Page and line numbers identify the revision locations in the original submission manuscript. Green text shows the
 550 targeted revisions. Black text provides surrounding context.

551 Section 2.4.5. Cropping Intensity Estimation

552 We estimated single-cropping and double-cropping within mapped rice paddies by identifying distinct
 553 periods of NDVI increase and decline. Following previous cropping-intensity studies, we used peak
 554 height, temporal separation, prominence, and duration to distinguish sustained growth periods from
 555 small fluctuations in the fitted curves (C. Liu et al., 2020; Guo et al., 2021; Ren et al., 2024). We
 556 combined these literature-based criteria with additional observation checks and selected the final rule
 557 using 343 manually interpreted locations. Table S1 lists the numerical conditions and their sources or
 558 adaptations.

559 The second-order harmonic curve provided the initial cropping-intensity estimate. Because this curve
 560 could merge two separate growth periods, we also examined local NDVI patterns using second-order
 561 Savitzky–Golay smoothing with three window lengths. We applied these windows to daily linearly
 562 interpolated NDVI to examine growth patterns at different smoothing scales. A local cycle had to show
 563 a sufficiently high peak, a clear increase and decline, and sufficient duration. Adjacent cycles also had
 564 to be separated in time.

565 We checked the local cycles against the original NDVI observations to avoid accepting peaks supported
566 only by interpolation or smoothing. Each cycle required observations near the peak and both adjacent
567 troughs, and the observed values had to support the increase and decline. Figure 4 illustrates the seasonal
568 patterns associated with single-cropping and double-cropping. We masked the resulting maps using the
569 final rice paddy maps and calculated regional double-cropping rates as the double-cropped area divided
570 by the area classified as single-cropping or double-cropping.

571 Figure 4 shows an example HLS time series with one or two distinct growth periods within mapped rice
572 paddies. Before vegetation growth, LSWI exceeded NDVI, indicating surface water consistent with rice
573 transplanting. These patterns support the identification of potential single-cropping and double-
574 cropping within rice paddies, but do not establish that every growth period represents rice cultivation.
575 Double-cropping can include rice followed by another crop, which would not require a second flooding
576 signal.

577 **Section 3.4. Regional rice paddy density and cropping intensity**

578 For cropping intensity, the proposed estimation method has achieved an overall classification agreement
579 of 89.8 % across the 343 manually interpreted locations used for rule selection, comprising 186 single-
580 cropping and 157 double-cropping locations. Agreement has reached 92.5 % for single-cropping and
581 86.6 % for double-cropping.

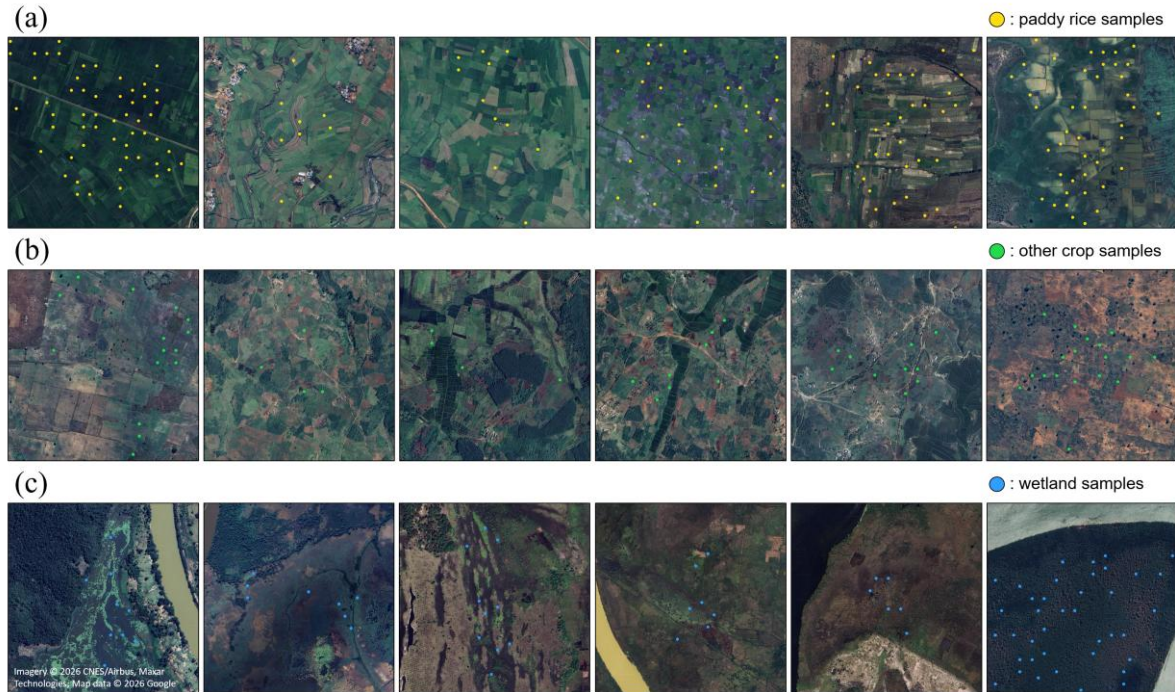
582

583 **Major Comment 2.**

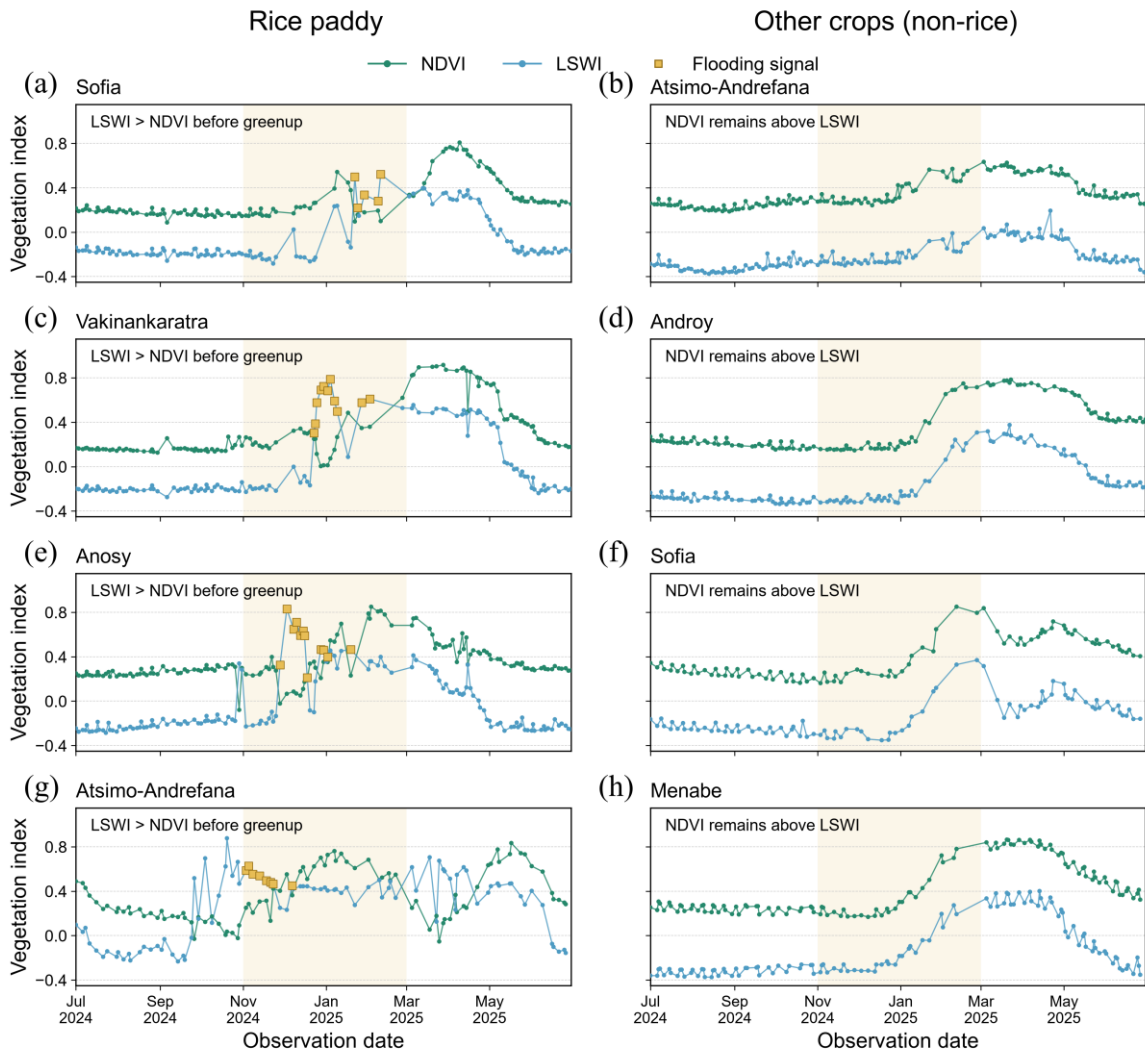
584 *The validation dataset described in Section 2.3.4 requires substantially more detail. For a data product*
585 *paper, the construction of the reference dataset is a central part of the study and should be sufficiently*
586 *transparent for readers to judge its reliability and reproducibility. The manuscript currently states that*
587 *samples were manually interpreted using Sentinel-2 RGB imagery, HLS-derived NDVI and LSWI time*
588 *series, and local expertise, but the actual labeling protocol remains unclear. For example, what specific*
589 *visual or temporal criteria were used to label a sample as rice or non-rice? How were NDVI and LSWI*
590 *used together during the interpretation process? These details are necessary, particularly because the*
591 *reported accuracy is one of the major strengths claimed for the dataset.*

592 *Response: Thank you very much for highlighting the importance of a clear reference-labeling protocol.*
593 *We have expanded Section 2.3.4 to explain the visual and temporal criteria used to distinguish rice*
594 *paddies from other crops and how we considered the imagery and both indices together. We have also*
595 *clarified the high-resolution imagery source to Google Earth, added reference time-series examples,*
596 *and made the reference dataset publicly available on Zenodo.*

597 *We identified agricultural land from clearly defined field boundaries that formed repeated patterns*
598 *across cultivated areas in high-resolution Google Earth RGB imagery. We investigated homogeneous*
599 *field interiors and avoided ambiguous boundaries where a pixel could include crops and surrounding*
600 *land covers. Within these areas, dark-green vegetation and a wet appearance supported rice-paddy*
601 *interpretation, whereas other-crop fields generally appeared drier. The original Figure S2 illustrates*
602 *these contrasting field appearances using pseudo-labeled samples. We have shown this figure below as*
603 *visual context and have provided separate examples from the reference dataset in the time-series figure*
604 *below.*



605
606 **Figure S2. Zoomed-in visualizations of pseudo-labeled samples showing their spatial distribution generated using**
607 **phenology-based rules for (a) rice paddy, (b) other crops, and (c) wetlands across diverse regions of Madagascar**
608 **(background imagery: Imagery © 2025 CNES/Airbus, Maxar Technologies; Map data © 2026 Google).**



609

610 Figure R4. HLS-derived NDVI and LSWI time series at four rice-paddy reference locations (a, c, e, g) and four
 611 other-crop reference locations (b, d, f, h). Green and blue indicate NDVI and LSWI, respectively. Shaded areas
 612 indicate the typical November–February transplanting period, and yellow squares mark observations with LSWI
 613 above NDVI before crop growth.

614 We used the seasonal relationship between NDVI and LSWI to assess whether the visual evidence was
 615 consistent with rice cultivation. For rice paddies, LSWI temporarily exceeded NDVI during
 616 transplanting, providing evidence of surface water before substantial canopy development. A
 617 subsequent increase in NDVI supported the transition from flooding to crop growth. For other crops,
 618 NDVI remained clearly higher than LSWI through the annual observation period, together with a drier
 619 appearance in the imagery. Thus, rice labels required consistent visual and temporal evidence, including
 620 the flooding-to-greenup sequence, rather than an isolated crossing of the two indices. The figure above
 621 shows these patterns at four rice-paddy and four other-crop reference locations.

622 *Hervet Randriamady*, a co-author and local expert from Madagascar, interpreted the reference samples
 623 separately from training refinement. He did not participate in hard-sample interpretation and prepared
 624 the reference labels without consulting the final rice mapping products. The reference samples were not
 625 used for model training, and validation results did not guide hard-sample selection. Finally, the dataset
 626 contained 1,000 rice-paddy and 1,000 other-crop locations. We have made all 2,000 reference locations
 627 and their class labels available in CSV, GeoJSON and GeoPackage formats on Zenodo

(<https://doi.org/10.5281/zenodo.23292193>). Together, the expanded protocol, visual examples and public reference data allow readers to understand the labeling decisions and inspect the samples used for the accuracy assessment.

Revised manuscript

Page and line numbers identify the revision locations in the original submission manuscript. Green text shows the targeted revisions. Black text provides surrounding context.

Section 2.3.4. Validation Datasets

[Page 9, Line 22] We evaluated the land cover classification and rice paddy detection results using independently generated reference samples. For rice paddy detection, a local expert from Madagascar interpreted the rice-paddy and other-crop reference samples using high-resolution Google Earth RGB imagery, HLS-derived NDVI and LSWI time series, and local knowledge of rice cultivation. Clearly defined field boundaries and repeated field patterns supported identification of agricultural land, and we examined homogeneous interiors to avoid mixed boundaries. Rice-paddy interpretation combined dark-green, wet-looking fields with LSWI temporarily exceeding NDVI during transplanting and a subsequent increase in NDVI. Other-crop interpretation combined drier cultivated fields with NDVI remaining clearly above LSWI through the annual observation period. We considered the imagery and the seasonal relationship between both indices together when assigning labels.

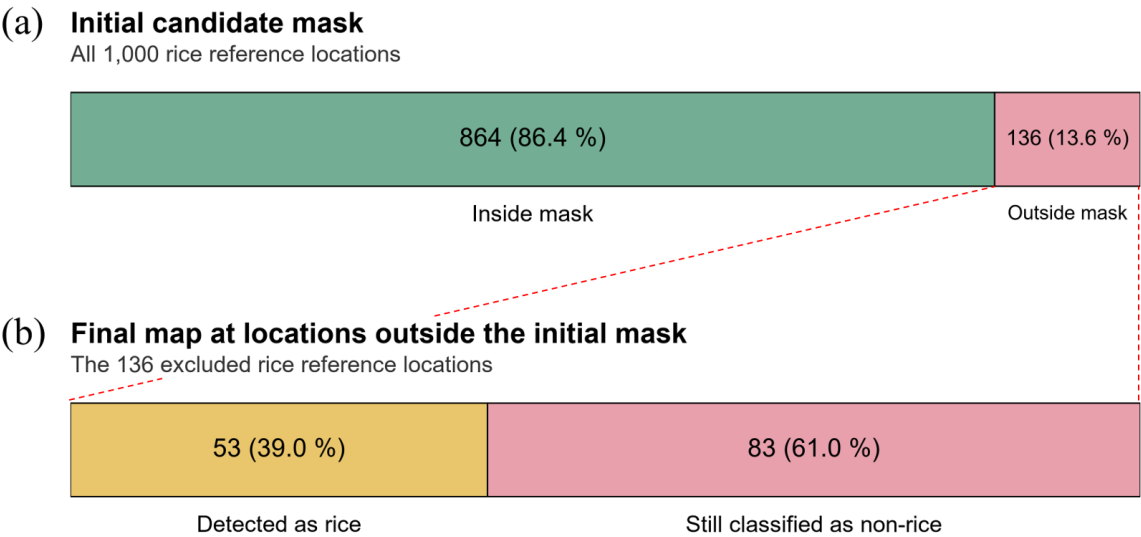
[Page 9, Line 22] Additionally, the reference interpretation and training refinement were separate processes. The reference labels were prepared after map production. The reference interpreter did not participate in hard-sample interpretation or consult the final rice map. We did not use reference samples for training or validation results to select hard samples. After removing duplicate coordinates, the reference dataset contained 1,000 rice-paddy and 1,000 other-crop locations for the assessment. We have made the reference coordinates and class labels publicly available at <https://doi.org/10.5281/zenodo.23292193>.

Major Comment 3.

I have a significant concern regarding the initial candidate-area extraction step. The workflow first trains a land-cover classifier using Dynamic World-derived samples and GSE features, and only pixels classified as cropland or flooded vegetation are retained for subsequent rice mapping. This creates a potentially important source of irreversible omission error. If a true rice pixel is classified as grass, shrub, water, or another land-cover class during this first step, it can never be recovered by the subsequent rice classifier. Although Section 2.3.4 states that the land-cover classification was evaluated using independent reference samples, what matters for this masking step is specifically its recall for true rice pixels. I therefore suggest reporting the proportion of independent rice reference samples that fall within the cropland/flooded-vegetation candidate mask. This would directly quantify the omission error introduced before the rice classifier is applied.

Response: Thank you very much for highlighting the risk of rice omission during candidate-area extraction. We agree that the recall of the candidate mask for rice is more informative for this purpose than overall land-cover accuracy. The candidate mask was designed to identify croplands and flooded vegetation, including wetlands that can be confused with rice paddies, and to focus the phenological

668 rules on these areas when generating high-confidence pseudo-labels. It also restricted the first-stage
669 output. However, it did not define the area that the final classifier could identify as rice.
670 We have evaluated the candidate mask by overlaying the 1,000 unique rice reference locations on the
671 mask used for the first-stage classification and checking whether each location fell inside or outside it.
672 The mask included 864 locations and excluded 136. Dividing the included count by all 1,000 rice
673 reference locations has yielded a rice recall of 86.4 %, with an initial exclusion rate of 13.6 % (Figure
674 S10a).



675
676 Figure S10. Initial candidate-mask coverage and subsequent rice detection across Madagascar. (a) Locations
677 inside and outside the candidate mask among 1,000 rice reference locations. (b) Final-map predictions for the 136
678 locations outside the mask.

679 The second-stage classifier could identify rice outside the initial candidate mask. We added other-land-
680 cover samples to the training data and classified rice, other crops, wetlands, and other land covers across
681 the country without applying the candidate mask again. The final map detected rice at 53 of the 136
682 excluded reference locations, recovering 39.0 % of this group (Figure S10b). The remaining 83
683 locations were still classified as non-rice. These results show that exclusion by the initial mask did not
684 necessarily lead to a rice omission in the final map, although the second stage did not recover every
685 excluded rice location. We have clarified these different roles of the candidate mask and the final
686 classification in the revised text for Sections 2.4.1 and 2.4.4.

687 Revised manuscript

688 Page and line numbers identify the revision locations in the original submission manuscript. Green text shows the
689 targeted revisions. Black text provides surrounding context.

690 Section 2.4.1. Candidate Pseudo-Labels Pools Extraction

691 [Page 10, Line 27] To define the candidate pools for rice paddy mapping, we first classified all 10 m
692 pixels into nine land cover classes following the Dynamic World scheme. We preserved cropland and
693 flooded vegetation as candidate areas for high-confidence pseudo-label generation, including wetlands
694 that could be confused with rice paddies. This mask restricted pseudo-label generation and the first-
695 stage output but did not define the prediction domain of the final classifier. Cropland pixels were treated
696 as potential rice paddy or other crops, while flooded vegetation pixels were treated as potential rice
697 paddy or wetlands.

698 Section 2.4.4. Rice Paddy Detection

699 [Page 13, Line 27] After the first-stage classification, we used the resulting confidence maps to identify
700 areas where phenology-based pseudo-labels were less sufficient, including fragmented field boundaries,
701 wetland margins, and spectrally mixed cropland. Visual inspection of the initial maps revealed errors
702 concentrated in low-confidence regions, including confusion between rice and wetlands and
703 misclassification of non-cropland areas. To address these errors, the authors, including a food system
704 expert from Madagascar, interpreted 36 polygons representing difficult rice, other-crop, and wetland
705 cases in low-confidence areas. For each year from 2017 to 2025, we reviewed Sentinel-2 RGB imagery
706 and Google Earth imagery, where available, to check for crop rotation, fallow periods, and land-use
707 changes before assigning the polygon labels. We generated 3,000 hard samples per year through
708 stratified random sampling within these polygons at the 10 m resolution of GSE, comprising 1,000
709 samples for each class. The sampled pixels inherited the labels assigned to their polygons for the
710 corresponding year and did not require individual interpretation. We combined these samples with the
711 pseudo-labels and other-land-cover samples automatically extracted through stratified random sampling
712 from the initial land cover map. A Random Forest classifier with 300 trees was then retrained to
713 distinguish rice, other crops, wetlands, and other land covers. We applied the final classifier across the
714 country without reapplying the initial candidate mask, allowing it to identify rice outside the candidate
715 areas.

716 Section 3.1. Accuracy assessment of rice paddy classification

717 [Page 17, Line 13] In contrast, the 2nd-stage classification missed 168 rice points and produced only 8
718 false rice predictions. The candidate-mask assessment has shown that 864 of the 1,000 rice reference
719 locations fell within the mask, giving a rice recall of 86.4 % (Fig. S10a). The final map detected rice at
720 53 of the 136 excluded locations, recovering 39.0 % of this group, while 83 locations remained
721 classified as non-rice (Fig. S10b). Initial exclusion therefore did not necessarily result in rice omission
722 in the final map. Across Köppen–Geiger climate zones, F1-scores ranged from 0.84 in the rainforest
723 zone to 0.96 in the steppe zone, with precision ranging from 0.98 to 1.00 (Fig. 6b).

724

725 **Major Comment 4.**

726 *The first-stage classifier is indeed trained using automatically generated pseudo-labels, but the final*
727 *product relies on manually interpreted hard samples from low-confidence areas. Therefore, I have some*
728 *concerns about repeatedly describing the framework as “near label-free.” In my view, this wording*
729 *somewhat overstates the degree to which the final mapping process is independent of manually labeled*
730 *data.*

731 Response: Thank you very much for highlighting the role of manual interpretation in the final mapping
732 framework. We agree that “near label-free” does not fully describe the two-stage approach. We have
733 therefore replaced this wording throughout the manuscript with “phenology-guided weak supervision
734 with uncertainty-guided refinement.” The first stage used phenological rules to generate pseudo-labels
735 automatically, while the second stage used model confidence to guide the interpretation of difficult areas.
736 This description makes the role of manual refinement explicit while preserving the contribution of
737 automatically generated pseudo-labels.

738 We have also clarified how manual interpretation contributed to second-stage training. The authors,
739 including a food system expert from Madagascar, interpreted 36 polygons representing difficult rice,
740 other-crop, and wetland cases in low-confidence areas. We then generated hard samples through
741 stratified random sampling within these polygons at the 10 m resolution of GSE. The sampled pixels
742 inherited the polygon labels and did not require separate manual interpretation each year. We combined
743 these samples with pseudo-labels and automatically sampled other-land-cover points to train the final
744 classifier.

745 The final-map accuracy therefore reflects the complete framework, including manual refinement. Our
746 comparison of training configurations has shown that pseudo-labels and targeted hard samples
747 contributed complementary information, with their combination achieving the highest accuracy (Fig.
748 S8). We have revised the Methods and Discussion to explain these roles and removed wording that
749 suggests the final maps were produced without manual interpretation.

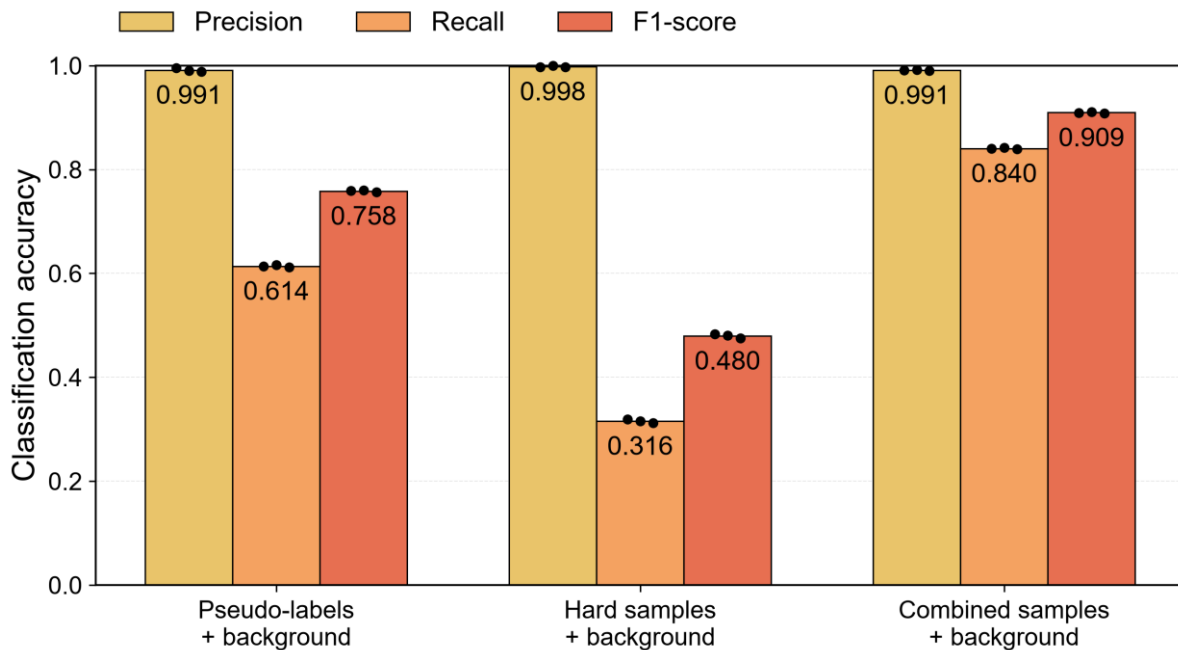


Figure S8. Accuracy of classifiers trained with pseudo-labels, hard samples, and their combination using the same GSE features and background samples. Bars and numerical labels show mean precision, recall, and F1-score across three random seeds, and black points indicate individual runs.

Revised manuscript

Page and line numbers identify the revision locations in the original submission manuscript. Green text shows the targeted revisions. Black text provides surrounding context.

Abstract

[Page 2, Line 10] Here, we present the first country-wide, annual rice paddy maps of Madagascar at 10 m resolution from 2017 to 2025. The two-stage Random Forest classification used GSE annual features within a framework of phenology-guided weak supervision with uncertainty-guided refinement. The first stage used automatically generated pseudo-labels, and the second stage incorporated samples extracted from manually interpreted polygons in low-confidence areas. The pseudo-labels formed compact and clearly separated clusters in the GSE feature space across all nine years, and only five GSE dimensions consistently contributed to rice discrimination, indicating that pre-trained embeddings encoded phenologically meaningful information for rice paddy.

[Page 2, Line 19] Our maps indicated an increase in mapped rice paddy extent from 886,112 ha in 2017 to 1,195,766 ha in 2025 (+34.9 %), with most gains occurring along the margins of existing paddies rather than in new frontiers. This study demonstrates that phenology-guided weak supervision with uncertainty-guided refinement provides a scalable approach for rice paddy mapping in data-scarce regions. The annual rice paddy maps are publicly available from Zenodo at <https://doi.org/10.5281/zenodo.20654510> (Kwon et al., 2026).

772 Section 1. Introduction

773 [Page 3, Line 28] Although there have been attempts to employ crowdsourced ground references to
774 produce rice paddy maps in Madagascar, spatially and temporally biased samples can substantially
775 constrain mapping accuracy in countries with heterogeneous landscapes (De Vos et al., 2023). In this
776 context, pseudo-labeling methods generate training samples from physically interpretable satellite
777 signals and provide weak supervision for mapping in data-scarce regions (You et al., 2023; Yang et al.,
778 2021). Nevertheless, the reliability of pseudo-labels has not been rigorously evaluated at the country
779 scale, particularly in ecologically heterogeneous environments such as Madagascar.

780 [Page 4, Line 10] These embeddings have therefore shown strong performance in general land cover
781 classification (Mantey et al., 2025; Li and Yu, 2025). Yet, their potential for crop type mapping,
782 particularly for rice paddy, has not been systematically evaluated. Integrating phenology-guided weak
783 supervision with GSE-based classification and uncertainty-guided refinement could provide a practical
784 rice mapping framework wherever both data sources are available.

785 [Page 4, Line 17] First, we generated pseudo-labels from optical satellite-derived phenological
786 indicators, using the LSWI–NDVI flooding signal to identify rice paddy without relying on extensive
787 ground reference data. Second, we classified GSE annual embeddings using these pseudo-labels and
788 used model confidence to guide the interpretation of difficult areas. We then refined the classifier with
789 samples extracted from the interpreted polygons, where fragmented fields, mixed pixels, or spectrally
790 similar land cover types can reduce classification reliability. Third, we applied harmonic NDVI fitting
791 to characterize cropping intensity over the mapped rice paddy areas.

792 Section 2.2. Overall Framework

793 [Page 5, Line 16] Our rice paddy mapping framework uses phenology-guided weak supervision with
794 uncertainty-guided refinement. It includes land cover classification, rice phenology analysis, rule-based
795 pseudo-label generation, two-stage Random Forest classification, and cropping intensity estimation
796 through harmonic NDVI fitting (Fig. 2). The rationale for separating pseudo-label generation (HLS, 30
797 m) from classification (GSE, 10 m) is that HLS provides the dense temporal observations needed to
798 detect rice-specific phenological signals, while GSE provides higher spatial resolution and gap-free
799 annual features for wall-to-wall mapping.

800 [Page 6, Line 4] To complement the clear rice signals captured by the pseudo-labels, the second
801 classification stage used model confidence to guide polygon interpretation in difficult areas. We
802 extracted hard samples from these polygons to refine the predictions.

803 Section 2.4.4. Rice Paddy Detection

804 [Page 13, Line 27] After the first-stage classification, we used the resulting confidence maps to identify
805 areas where phenology-based pseudo-labels were less sufficient, including fragmented field boundaries,
806 wetland margins, and spectrally mixed cropland. Visual inspection of the initial maps revealed errors
807 concentrated in low-confidence regions, including confusion between rice and wetlands and
808 misclassification of non-cropland areas. To address these errors, the authors, including a food system
809 expert from Madagascar, interpreted 36 polygons representing difficult rice, other-crop, and wetland
810 cases in low-confidence areas. For each year from 2017 to 2025, we reviewed Sentinel-2 RGB imagery
811 and Google Earth imagery, where available, to check for crop rotation, fallow periods, and land-use
812 changes before assigning the polygon labels. We generated 3,000 hard samples per year through
813 stratified random sampling within these polygons at the 10 m resolution of GSE, comprising 1,000
814 samples for each class. The sampled pixels inherited the labels assigned to their polygons for the
815 corresponding year and did not require individual interpretation. We combined these samples with the
816 pseudo-labels and other-land-cover samples automatically extracted through stratified random sampling
817 from the initial land cover map. A Random Forest classifier with 300 trees was then retrained to
818 distinguish rice, other crops, wetlands, and other land covers. We applied the final classifier across the
819 country without reapplying the initial candidate mask, allowing it to identify rice outside the candidate
820 areas.

821 Section 2.5. Accuracy Assessment

822 [Page 15, Line 13] To evaluate classification performance, we compared the predicted maps with the
823 reference samples described in Section 2.3.4. To investigate the contributions of the training samples,
824 we have compared models trained with pseudo-labels, hard samples, and their combination using the
825 same GSE features, background samples, Random Forest settings, and reference locations. The
826 background samples represented other land covers and were automatically extracted from the initial
827 land cover map. Overall accuracy (OA) was calculated as the proportion of correctly classified samples
828 among all validation samples.

829 Section 3.1. Accuracy assessment of rice paddy classification

830 [Page 17, Line 13] In contrast, the 2nd-stage classification missed 168 rice points and produced only 8
831 false rice predictions. The comparison of training configurations has further clarified the contribution
832 of each sample source, yielding F1-scores of 0.756–0.760 for pseudo-labels with background samples,
833 0.475–0.483 for hard samples with background samples, and 0.909–0.911 for their combination (Fig.
834 S8). The highest accuracy has resulted from combining the two sample sources, supporting their
835 complementary contributions to rice detection. Across Köppen–Geiger climate zones, F1-scores ranged
836 from 0.84 in the rainforest zone to 0.96 in the steppe zone, with precision ranging from 0.98 to 1.00
837 (Fig. 6b).

838 Section 4.2. Pre-trained embeddings captured rice phenology in a compact feature space

839 [Page 28, Line 17] Because GSE integrates information from multiple sensors and summarizes broader
840 spectral and temporal patterns, it provided a compact input that helped reduce reliance on cloud-gap
841 filling and manual feature engineering. Our targeted refinement helped the classifier use the information
842 in GSE embeddings to distinguish difficult rice and non-rice cases. Errors along field edges and wetland
843 margins in the first-stage map indicate that a compact feature space alone does not ensure accurate
844 classification where land covers share similar characteristics (Fig. S5). Adding samples from these
845 difficult areas improved classification while preserving the broader information supplied by pseudo-
846 labels (Fig. S8). This suggests that the usefulness of pre-trained embeddings depends on training
847 samples that represent both clear phenological patterns and ambiguous land-cover conditions.
848 Combining these sample sources therefore helped translate the information captured by GSE into more
849 reliable rice predictions without constructing additional crop-specific features.

850 Section 5. Conclusions

851 [Page 30, Line 14] This study demonstrates a scalable approach for annual rice paddy mapping in data-
852 scarce regions through phenology-guided weak supervision with uncertainty-guided refinement using
853 pre-trained satellite embeddings. Rather than relying on dense field reference data, the framework uses
854 the seasonal flooding-to-greenup sequence of transplanted rice to generate training samples and uses
855 GSE to represent rice-related phenology in a compact feature space.

856

857 Major Comment 5.

858 *I suggest revising Section 2.3.5, currently titled “Benchmark rice mapping product.” The Africa-wide*
859 *20 m rice product is useful for comparison, but I do not think it needs to be framed as a “benchmark”*
860 *dataset, particularly when the manuscript later emphasizes that the proposed product substantially*
861 *outperforms it on the authors’ validation samples.*

862 *Response:* Thank you very much for suggesting a clearer description of the comparison product. We
863 agree that the Africa-wide rice map is useful for comparison and that describing it as a “benchmark” is
864 unnecessary. We have therefore renamed Section 2.3.5 “Existing rice mapping products” and replaced
865 “benchmark” with “the rice map from Jiang et al. (2025)” or “the comparison product” throughout the
866 manuscript. These changes clarify that we used an existing rice map to compare classification
867 performance and spatial patterns with our results. We have also revised the wording around the
868 comparison to describe the observed differences between the maps on our reference dataset.

869 Revised manuscript

870 Page and line numbers identify the revision locations in the original submission manuscript. Green text shows the
871 targeted revisions. Black text provides surrounding context.

872 Section 2.3.5. Existing rice mapping products

873 [Page 10, Line 2] 2.3.5 Existing rice mapping products

874 [Page 10, Line 11] We compared our 2023 rice paddy map with the 2023 Africa-wide rice map produced
875 by Jiang et al. (2025). This product covers 34 rice-producing African countries, including Madagascar.
876 It combined Sentinel-2 optical information for object-based segmentation with Sentinel-1 SAR features
877 for Random Forest classification. We first evaluated both maps at their native resolutions, grouping
878 mapped classes as rice and non-rice. To account for their different spatial resolutions, we have also
879 compared both maps on common 20 m cells, retaining cells with at least 95 % valid coverage in each
880 product and identifying rice where it covered at least half of the cell.

881

882 **Major Comment 6.**

883 *Some parts of the Methods contain unnecessary textbook-style explanations that interrupt the scientific*
884 *flow. For example, the sentence stating that “The Random Forest algorithm is an ensemble technique*
885 *that produces predictions by aggregating the outputs of multiple decision trees for classification or*
886 *regression tasks” is not necessary for an ESSD readership. Random Forest is a well-established method,*
887 *and a citation is sufficient.*

888 *Response: Thank you very much for recommending a more concise Methods section. We agree that a*
889 *general explanation of Random Forest is unnecessary for the ESSD readership and interrupts the*
890 *description of our workflow. We have therefore removed the quoted sentence and retained the citation*
891 *to Breiman (2001) alongside the description of how we applied the classifier. The revised paragraph*
892 *focuses on the input features, training samples, and classifier settings used in this study. This change*
893 *keeps the Methods focused on the information needed to understand and reproduce our analysis.*

894 **Revised manuscript**

895 *Page and line numbers identify the revision locations in the original submission manuscript. Green text shows the*
896 *targeted revisions. Black text provides surrounding context.*

897 **Section 2.4.1. Candidate Pseudo-Labels Pools Extraction**

898 [Page 10, Line 23] From these stable pixels, we sampled up to 1,000 pixels per land cover class per
899 region, resulting in a stratified reference dataset without manual labeling. We then trained a Random
900 Forest classifier with 200 trees for each region and year using the GSE annual embeddings and the
901 Dynamic World-derived samples (Breiman, 2001). To define the candidate pools for rice paddy
902 mapping, we first classified all 10 m pixels into nine land cover classes following the Dynamic World
903 scheme.

904

905 **Community comments**

906 *An interesting work that nicely reflects the recent developments in national crop type/rice mapping*
907 *research. While I agree with most of the findings of the authors, it would be interesting to further*
908 *investigate how the authors methodological choices indeed improve on existing national-scale maps.*

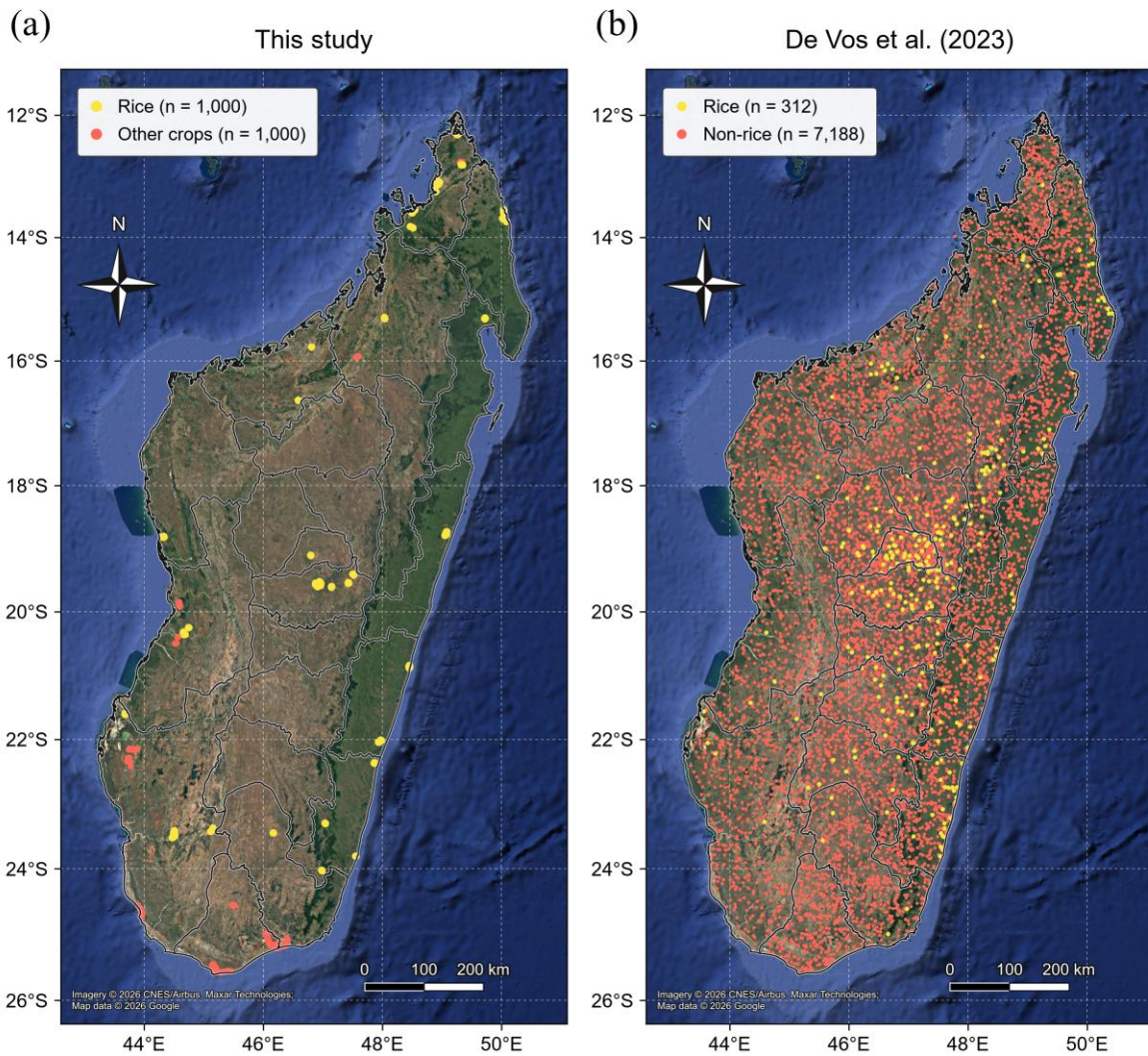
909 *Response: Thank you very much for your positive assessment of our study and your constructive*
910 *suggestion to further evaluate how our methodological frameworks improve rice paddy mapping at the*
911 *national scale.*

912

913 **Major Comment 1.**

914 *A comparison with the continental rice distribution map shows clear improvement, but this is expected*
915 *upfront because the difference in scale (continental vs. national) and thereby representation of data*
916 *used for training. The authors identify other existing national-scale maps such as the one we made a*
917 *couple of years ago (De Vos et al. (2023)). The latter map is definitely biased towards very clear rice*
918 *paddies and is missing out on smaller and further-from-the-road fields because of parts of its sampling*
919 *design (particularly the ones only using crowdsourced data). It would be very informative to understand*
920 *how this mapping model compares to our earlier maps and whether this study could highlight for which*
921 *areas/regions the improvements are most visible. It would therefore be interesting to see the training*
922 *samples and/or national-scale maps to be used as additional benchmark. Feel free to reach out to gain*
923 *access to our training samples and maps.*

924 *Response: Thank you very much for the constructive comments and for kindly sharing the national map*
925 *and independent test data. We agree that comparison with a continental product alone cannot establish*
926 *the benefits of our methodological choices. We have now compared both national products using the*
927 *reference data provided by De Vos et al. (2023) and our manually interpreted reference data, and added*
928 *brief descriptions to Sections 2.3.5, 2.5, 3.1, and 4.1, with the figures in the revised Supplement (Figures*
929 *S1, S6 and S3).*

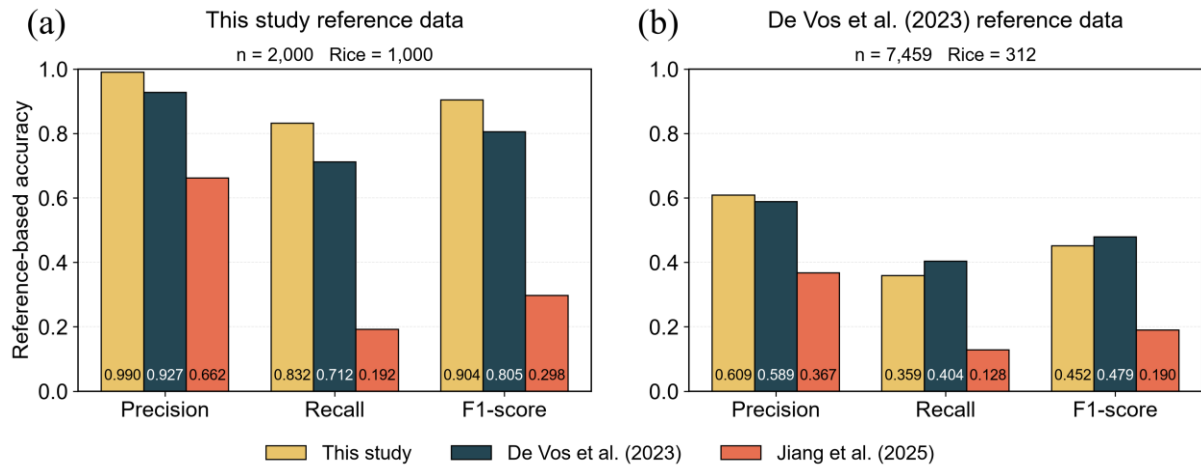


930

931 Figure S1. Spatial distribution of reference samples in Madagascar from (a) this study and (b) De Vos et al. (2023).
 932 Yellow points indicate rice, and red points indicate other crops in (a) and non-rice in (b) (background imagery:
 933 Imagery © 2026 CNES/Airbus, Maxar Technologies; Map data © 2026 Google).

934 We first clarified the temporal and thematic scope of the two products. The rice paddy maps developed
 935 by De Vos et al. (2023) represent rice cultivation using multi-annual Sentinel-1 and Sentinel-2
 936 observations from 2017 to 2020, rather than a single-year map. We therefore compared our 2017–2020
 937 annual maps and investigated temporarily aggregated conditions. As in our study, De Vos et al. (2023)
 938 explicitly excluded upland rice because it lacks a flooding stage. Both studies consequently focus on
 939 paddy systems, and our phenological rules target the flooding-to-greenup sequence of transplanted rice.

940 First, our rice paddy map showed higher accuracy against our rice and other-crop reference dataset. At
 941 the 2,000 sampled locations, our map achieved precision of 99.0 %, recall of 83.2 %, and an F1-score
 942 of 0.904, compared with 92.7 %, 71.2 %, and 0.805, respectively, for the rice paddy map produced by
 943 De Vos et al. (2023) (Figure S6a). Our map detected 186 rice locations missed by their map but missed
 944 66 locations that their map detected. Our map maintained a higher F1-score than the map from De Vos
 945 et al. (2023) in the 10 km spatial-block sensitivity analysis, with a conditional 95 % interval of 0.053–
 946 0.172 for the difference. This result supports improved separation of rice and other crops in our
 947 reference dataset.

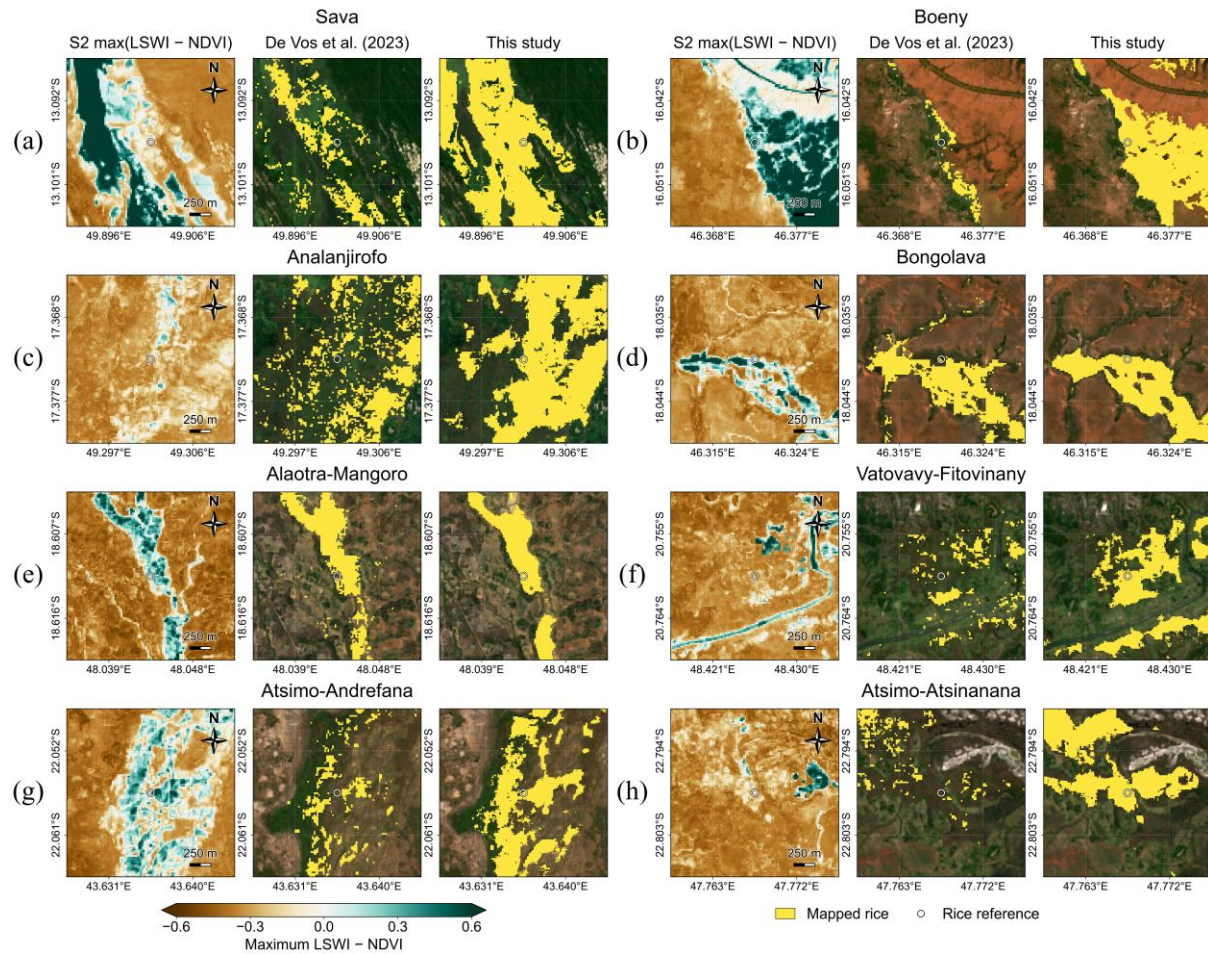


948

949 Figure S6. Precision, recall, and F1-score of rice maps from this study, De Vos et al. (2023), and Jiang et al. (2025),
 950 evaluated using reference data from (a) this study and (b) De Vos et al. (2023).

951 Second, the reference data from De Vos et al. (2023) provide a complementary assessment of our multi-
 952 annual rice extent. To compare the maps over the same nominal period, we combined our annual rice
 953 maps from 2017 to 2020, classifying a location as rice if at least one annual map identified it as rice.
 954 We evaluated both maps at 7,459 locations with common coverage, including all 312 rice reference
 955 locations. Our combined map achieved a precision of 60.9 %, slightly exceeding the precision of 58.9 %
 956 obtained for the map produced by De Vos et al. (2023). However, our combined map achieved a recall
 957 of 35.9 %, which was lower than the recall of 40.4 % obtained for their map. The resulting F1-score
 958 was 0.452 for our combined map, compared with 0.479 for their map (Figure S6b). This additional
 959 national-scale comparison shows that our combined map produced fewer false rice detections on the
 960 shared reference dataset, although it also missed more rice locations than the map from De Vos et al.
 961 (2023).

962 Additionally, we have added spatial comparisons at several representative sites to exhibit where our rice
 963 paddy maps detected rice locations missed by De Vos et al. (2023) (Figure S3). At each selected
 964 reference location, our map classified the location as rice, consistent with the reference label provided
 965 by De Vos et al. (2023), whereas their map classified it as non-rice. These differences are particularly
 966 visible in Boeny, Vatovavy-Fitovinany, and Atsimo-Atsinanana, where our map shows more continuous
 967 rice patches around the reference locations (Figure S3b, f, and h). These comparisons have identified
 968 specific areas where our map reduced rice omissions.



969

970 Figure S3. Spatial comparison of rice maps at selected sites in (a) Sava, (b) Boeny, (c) Analanjirofo, (d) Bongolava,
 971 (e) Alaotra-Mangoro, (f) Vatovavy-Fitovinany, (g) Atsimo-Andrefana, and (h) Atsimo-Atsinanana. Left panels
 972 show maximum LSWI – NDVI from Sentinel-2 observations. Middle and right panels show rice paddy maps
 973 from De Vos et al. (2023) and this study, respectively. Yellow indicates mapped rice, and circles indicate rice
 974 reference samples from De Vos et al. (2023).

975 We have also investigated where our combined map missed rice reference locations to place these local
 976 improvements in the context of the overall comparison. Of the 312 rice reference locations provided by
 977 De Vos et al. (2023), our 2017–2020 combined map missed 200. Their map detected 46 of these
 978 locations, while both maps missed the remaining 154. The largest numbers of omissions in our map
 979 occurred in Itasy, Vakinankaratra, and Analamanga, with 30, 29, and 17 missed rice locations,
 980 respectively. We have added spatial examples from the eight regions containing the most missed
 981 locations, which together account for 148 of the 200 omissions. The figure below shows four locations
 982 detected only by De Vos et al. (2023) and four locations missed by both maps.

983 These examples show that missed reference locations do not all have the same observed flooding signal.
 984 Among the eight central reference locations, the maximum LSWI – NDVI from clear Sentinel-2
 985 observations in 2020 was positive at Analamanga and Itasy and remained at or below zero at the other
 986 six locations. The latter observations suggest that a weak or unobserved flooding signal may contribute
 987 to difficulty in identifying some rice paddies. However, omissions at the two locations with positive
 988 signals show that flooding evidence alone did not ensure rice detection by the final classifier. These
 989 images provide local evidence for interpreting the omissions, but they cannot establish their causes
 990 across the full 2017–2020 period. Together with Figure S3, this comparison shows both local

improvements and remaining omissions, while the lower recall on the external reference dataset highlights the need to improve rice detection across a wider range of paddy conditions.

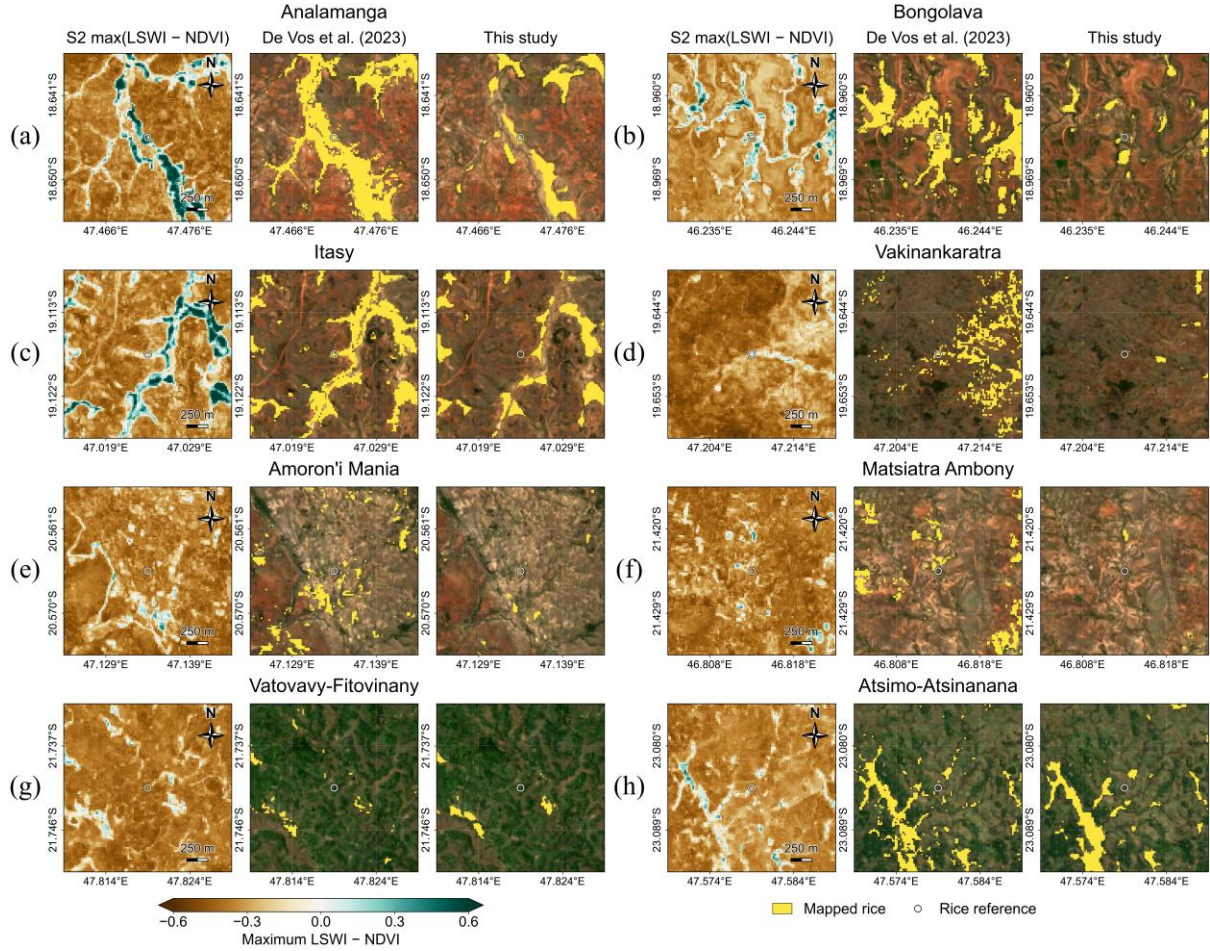


Figure R5. Spatial comparison at rice reference locations missed by our 2017–2020 combined map in (a) Analamanga, (b) Bongolava, (c) Itasy, (d) Vakinankaratra, (e) Amoron'i Mania, (f) Matsiatra Ambony, (g) Vatovavy-Fitovinany, (h) Atsimo-Atsinanana. Left panels show maximum LSWI – NDVI from Sentinel-2 observations in 2020. Middle and right panels show mapped rice from De Vos et al. (2023) and the union of our 2017–2020 annual maps, respectively, over the same 2020 Sentinel-2 RGB composites. Yellow indicates mapped rice, and circles indicate rice reference samples from De Vos et al. (2023). The central reference location is detected only by De Vos et al. (2023) in (a–d) and missed by both maps in (e–h).

To understand the difference between the two assessments, we have investigated all reference locations. Rice accounts for 50 % of our reference samples but only 4.2 % of the external samples, whose non-rice class also includes a wider range of land covers. We have recalculated the external F1-score assuming a 50 % rice proportion while keeping recall and the non-rice false-positive rate unchanged. This calculation has increased F1 from 0.452 to 0.524, while recall has remained 35.9 % (panel c of the figure below). Class proportions therefore affect F1 but do not explain the low rice recall.

We have also evaluated the same 2017–2020 combined map against both reference datasets. It has achieved F1-scores of 0.958 and 0.452, with recalls of 92.3 % and 35.9 %, on our and the external reference datasets, respectively (panel a of the figure below). The performance difference therefore remains when the map is held constant.

The reference datasets also represent different terrain and geographic conditions. We have found that 96.0 % of our rice locations lie in terrain below 5°, compared with 64.4 % of the external locations (Figure R6b). Our rice samples cover 14 regions, compared with 21 externally, and 88.0 % of our other-

crop samples are concentrated in Androy, Atsimo-Andrefana, and Anosy. Within the external dataset, our combined map has achieved recall of 50.7 % below 5° and 9.0 % at 5° or greater. Omissions are therefore more frequent in sloping landscapes, although surrounding terrain slope alone does not establish their cause.

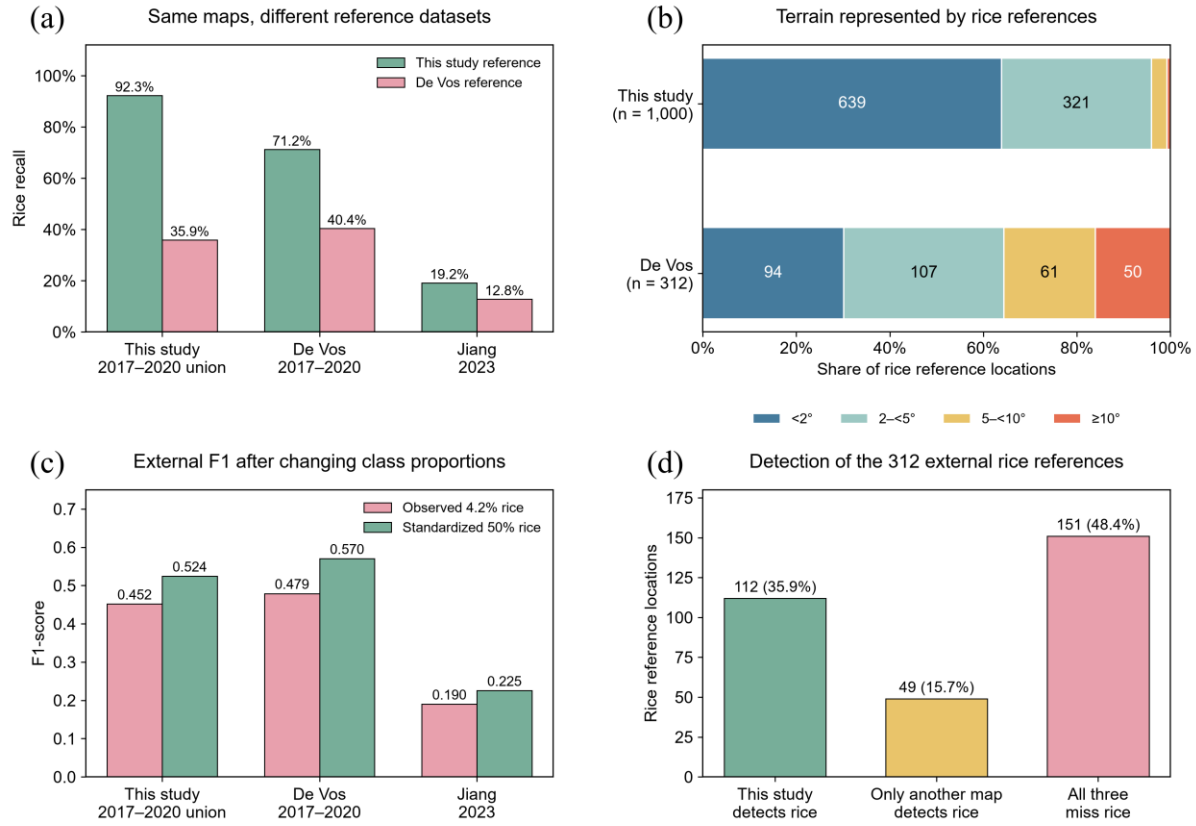
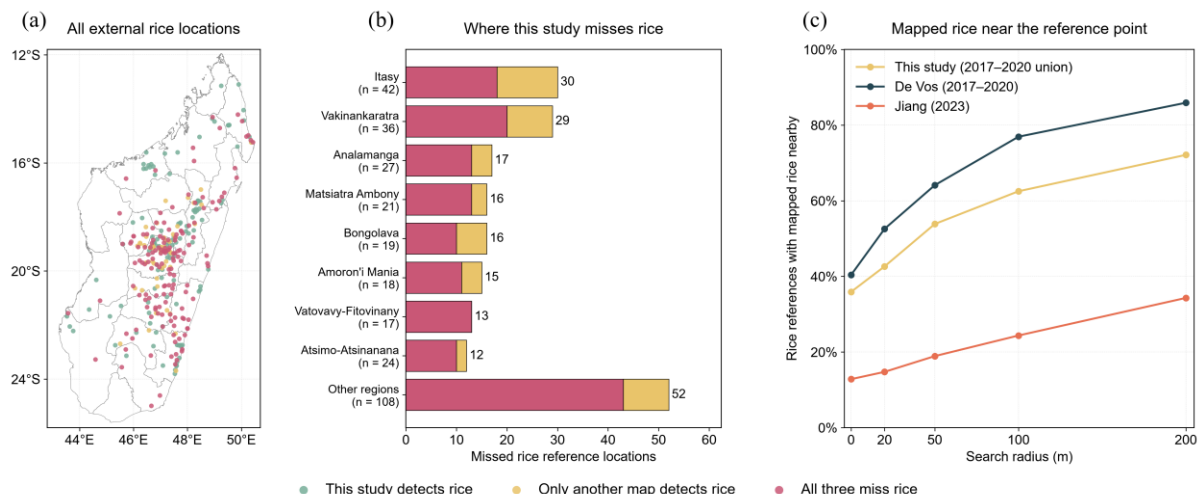


Figure R6. Reference composition and rice detection. (a) Recall of the same three maps on both reference datasets, using our 2017–2020 combined map. (b) Rice reference locations grouped by surrounding terrain slope. (c) External F1-scores using the observed rice proportion and a 50 % rice proportion, with recall and false-positive rates unchanged. (d) Detection of the 312 external rice locations by our map and the two comparison products.

We have also examined shared omissions and nearby mapped rice. Of the 200 rice locations missed by our combined map, 151 have also been missed by both comparison products, while 49 have been detected by at least one of them. These omissions occur across several regions, with the largest numbers in Itasy and Vakinankaratra (Figure R7a and b). Among our missed locations, 56 have mapped rice within 50 m, but 87 have none within 200 m. Field boundaries or positional differences may affect some comparisons, but cannot explain all omissions. Because nearby rice may belong to a different field, these distances describe spatial patterns rather than corrected accuracy (Figure R7c).



1030

1031 Figure R7. Locations and surroundings of the 312 external rice references. (a) Locations detected by our map,
 1032 detected only by the comparison products, or missed by all three. (b) Our missed locations by region, with n
 1033 showing the total rice reference count. (c) Proportion of reference locations with mapped rice within each search
 1034 radius. Distance is zero for locations mapped as rice and otherwise measured to the nearest rice pixel center.

1035 In summary, these comparisons show improved rice and other-crop separation in our reference dataset
 1036 and additional rice detections at several sites. The external assessment shows that these gains do not
 1037 extend equally across all sampled paddy conditions, particularly in more sloping landscapes. Broader
 1038 reference coverage and targeted refinement in these settings would help improve detection beyond the
 1039 conditions represented by our own samples.

1040 We appreciate Dr. De Vos again for kindly sharing the national rice map and independent reference data,
 1041 which have enabled an in-depth comparison of the two mapping products.

1042

1043 Revised manuscript

1044 Page and line numbers identify the revision locations in the original submission manuscript. Green text shows the
 1045 targeted revisions. Black text provides surrounding context.

1046 Section 2.3.5. Existing rice mapping products

1047 [Page 10, Line 13] We compared our 2023 rice paddy map with the 2023 Africa-wide rice map produced
 1048 by Jiang et al. (2025). To extend the comparison to an existing national product, we have also compared
 1049 our maps with the 20 m national rice map from De Vos et al. (2023), which combined Sentinel-1 and
 1050 Sentinel-2 observations from 2017 to 2020 and excluded upland rice without a flooding stage. De Vos
 1051 et al. (2023) also provided independent rice and non-rice reference data, which we have used alongside
 1052 our own reference dataset to compare the two maps (Fig. S1).

1053 Section 2.5. Accuracy Assessment

1054 [Page 15, Line 13] To evaluate classification performance, we compared the predicted maps with the
 1055 reference samples described in Section 2.3.4. For the national comparison with De Vos et al. (2023), we
 1056 have evaluated both maps at the same locations within each reference dataset, using 2,000 locations
 1057 from this study and 7,459 locations from De Vos et al. (2023). For the latter comparison, we have
 1058 combined our 2017–2020 maps by classifying a location as rice if at least one annual map identifies it
 1059 as rice. We have also assessed the F1-score difference on our reference dataset by resampling paired

1060 predictions within 10 km spatial blocks. Overall accuracy (OA) was calculated as the proportion of
1061 correctly classified samples among all validation samples.

1062 Section 3.1. Accuracy assessment of rice paddy classification

1063 [Page 15, Line 24] The classification captured diverse paddy configurations, such as irrigated clusters
1064 on floodplains, fragmented plots along riparian corridors, and smallholder paddies embedded within
1065 mixed cropland mosaics. Comparison with De Vos et al. (2023) has also identified local improvements
1066 in rice detection, with our rice paddy maps detecting rice reference locations missed by their map (Fig.
1067 S3). These differences are particularly visible at sites in Boeny, Vatovavy-Fitovinany, and Atsimo-
1068 Atsinanana, where our map shows more continuous rice patches around the reference locations. At the
1069 10 m spatial resolution, paddy boundaries were generally well delineated, while adjacent land-cover
1070 types such as roads, bare soil, forest edges, and urban areas were largely excluded from the rice paddy
1071 class.

1072 [Page 17, Line 13] In contrast, the 2nd-stage classification missed 168 rice points and produced only 8
1073 false rice predictions. The national comparison has provided a further assessment using both reference
1074 datasets. On our reference dataset, it has yielded an F1-score of 0.904 for our map and 0.805 for De Vos
1075 et al. (2023) (Fig. S6a), with a conditional 95 % interval of 0.053–0.172 for the difference under 10 km
1076 spatial-block resampling. On the reference dataset provided by De Vos et al. (2023), our combined
1077 2017–2020 map has achieved a precision of 60.9 %, exceeding the 58.9 % obtained for their map.
1078 However, our combined map has achieved a lower recall of 35.9 %, compared with 40.4 % for their
1079 map, resulting in F1-scores of 0.452 and 0.479, respectively (Fig. S6b). Across Köppen–Geiger climate
1080 zones, F1-scores ranged from 0.84 in the rainforest zone to 0.96 in the steppe zone, with precision
1081 ranging from 0.98 to 1.00 (Fig. 6b).

1082 Section 4.1. Phenology-guided pseudo-labels supported annual rice mapping

1083 [Page 27, Line 14] Annual percentile thresholds allowed the rules to adjust to each year's conditions
1084 without prior knowledge of local rice calendars, supporting consistent sample generation across
1085 Madagascar. The national comparison demonstrated improved rice detection with less confusion
1086 between rice and other crops in our reference dataset. Higher precision, recall, and F1-score relative to
1087 De Vos et al. (2023) showed that our map detected more rice locations while producing fewer false rice
1088 classifications (Fig. S6a). More continuous rice patches around reference locations in Boeny, Vatovavy-
1089 Fitovinany, and Atsimo-Atsinanana illustrated these local improvements (Fig. S3). These findings
1090 support combining phenological samples with targeted refinement where a single training source may
1091 not adequately represent local conditions. Higher precision but lower recall on the external reference
1092 dataset suggests that broader coverage of rice-growing conditions could help address remaining
1093 omissions (Fig. S6b).

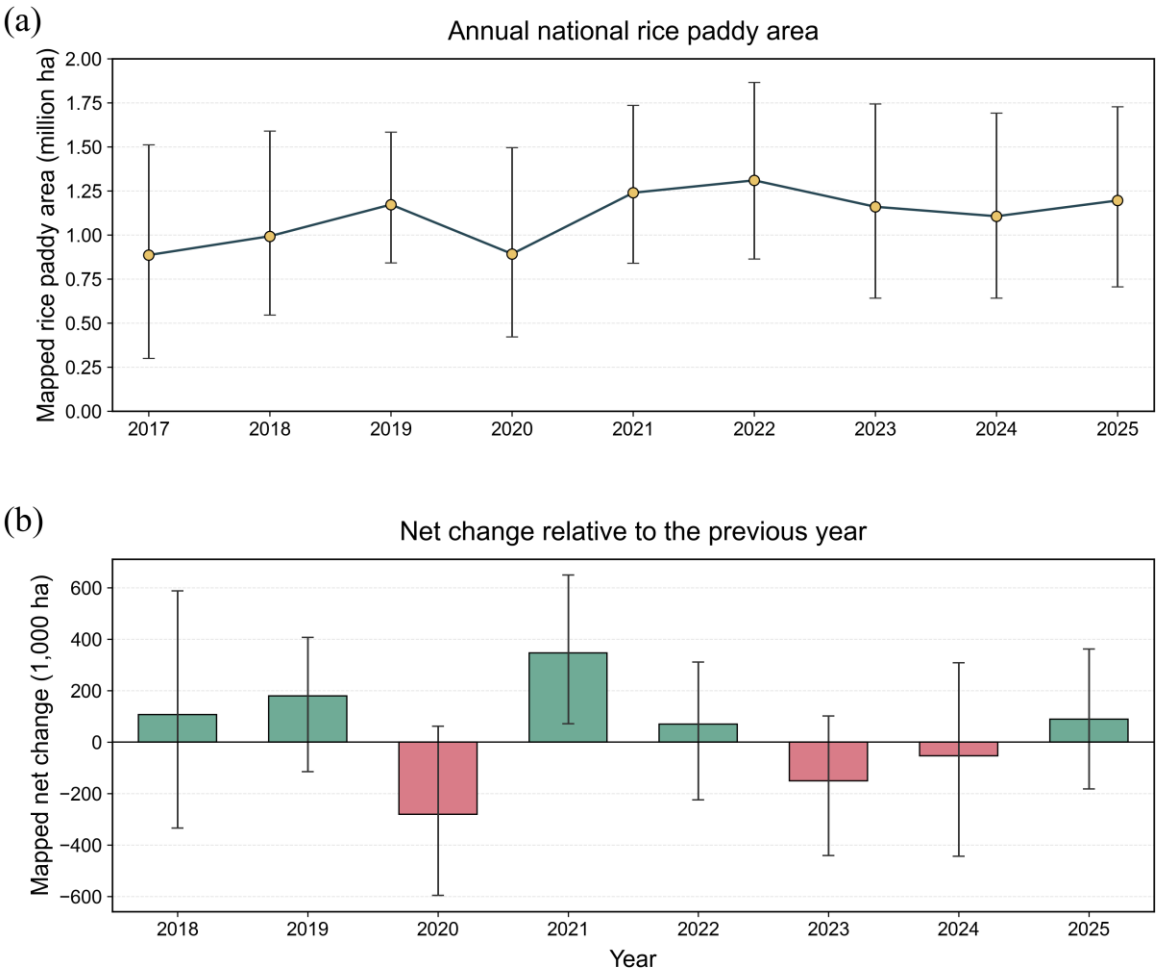
1094

1095 Major Comment 2.

1096 *Other than that, I would advise the authors to consider uncertainty estimates on the regional area*
1097 *calculations. Particularly the best practices outlined by Olofsson et al. (2014) are very informative on*
1098 *this and could aid framing the observed changes in a correct and uncertainty-aware way.*

1099 *Response:* Thank you very much for recommending the good-practice framework of Olofsson et al.
1100 (2014). We agree that uncertainty estimates help interpret rice paddy area and its changes. Following
1101 the suggestion, we have expanded the available sample set and conducted an additional national-scale
1102 analysis of year-specific uncertainty during 2017–2025.

1103 We have combined our 1,000 rice and 1,000 other-crop reference samples with 7,000 other-land-cover
 1104 samples for each year. These additional samples were drawn from the Dynamic World-derived samples
 1105 used for the initial land-cover classification, excluding locations that shared a mapped pixel with the
 1106 final rice-classifier training samples or our reference samples. We have preserved the original rice and
 1107 other-crop labels across all target years and treated the additional land-cover labels as non-rice.
 1108 For each year, we have used the corresponding map predictions and ground-area weights to calculate
 1109 conditional area estimates using the area-weighted error-matrix approach of Olofsson et al. (2014). We
 1110 have considered mapped rice, other crops, wetlands, and other land covers separately, rather than
 1111 applying one crop-derived rice proportion to the entire mapped non-rice area. We have then used 3,000
 1112 paired resamples of 10 km spatial blocks to calculate year-specific uncertainty widths. For net change,
 1113 we have calculated the difference between consecutive years within each resample, accounting for the
 1114 shared reference locations across years. The figure below retains the original mapped pixel-area totals
 1115 and shows the calculated uncertainty widths around these values. Panel (a) shows annual mapped area,
 1116 and panel (b) shows positive and negative changes relative to the previous year.



1117
 1118 Figure R8. Annual mapped rice paddy area and change in Madagascar during 2017–2025. (a) Original annual
 1119 mapped areas. (b) Net change relative to the previous year. Bars show year-specific uncertainty widths transferred
 1120 from the expanded-sample analysis using area-weighted estimates and paired 10 km spatial-block resampling.
 1121 The change calculations account for shared reference locations across years.

1122

1123 References

- 1124 Breiman, L.: Random forests, *Mach. Learn.*, 45, 5-32, <https://doi.org/10.1023/A:1010933404324>, 2001.
- 1125 De Vos, K., Campforts, B., Tits, L., Van Tricht, K., Bonte, K., Govers, G., and Jacobs, L.: Potential and limitations
1126 of crowdsourced data for high-resolution rice mapping in Madagascar: The importance of representation, *Int.*
1127 *J. Appl. Earth Obs. Geoinf.*, 117, 103204, <https://doi.org/10.1016/j.jag.2023.103204>, 2023.
- 1128 Guo, Y., Xia, H., Pan, L., Zhao, X., Li, R., Bian, X., Wang, R., and Yu, C.: Development of a New Phenology
1129 Algorithm for Fine Mapping of Cropping Intensity in Complex Planting Areas Using Sentinel-2 and Google
1130 Earth Engine, *ISPRS Int. J. Geo-Inf.*, 10, 587, <https://doi.org/10.3390/ijgi10090587>, 2021.
- 1131 Jiang, J., Zhang, H., Ge, J., Zuo, L., Xu, L., Song, M., Ding, Y., Xie, Y., and Huang, W.: The 20 m Africa rice
1132 distribution map of 2023, *Earth Syst. Sci. Data*, 17, 1781-1805, <https://doi.org/10.5194/essd-17-1781-2025>,
1133 2025.
- 1134 Kwon, R., Ryu, Y., Ciais, P., Makowski, D., Choi, C., Feng, H., Jeong, S., Tang, P., and Lee, K.: Enabling maize
1135 mapping in double-cropping regions of Argentina and Brazil without ground reference data by leveraging
1136 multiple satellite platforms, *Int. J. Appl. Earth Obs. Geoinf.*, 152, 105489,
1137 <https://doi.org/10.1016/j.jag.2026.105489>, 2026.
- 1138 Liu, C., Zhang, Q., Tao, S., Qi, J., Ding, M., Guan, Q., Wu, B., Zhang, M., Nabil, M., Tian, F., Zeng, H., Zhang,
1139 N., Bavuudorj, G., Rukundo, E., Liu, W., Bofana, J., Beyene, A. N., and Elnashar, A.: A new framework to map
1140 fine resolution cropping intensity across the globe: Algorithm, validation, and implication, *Remote Sens.*
1141 *Environ.*, 251, 112095, <https://doi.org/10.1016/j.rse.2020.112095>, 2020.
- 1142 Liu, L., Xiao, X., Qin, Y., Wang, J., Xu, X., Hu, Y., and Qiao, Z.: Mapping cropping intensity in China using time
1143 series Landsat and Sentinel-2 images and Google Earth Engine, *Remote Sens. Environ.*, 239, 111624,
1144 <https://doi.org/10.1016/j.rse.2019.111624>, 2020.
- 1145 Olofsson, P., Foody, G. M., Herold, M., Stehman, S. V., Woodcock, C. E., and Wulder, M. A.: Good practices for
1146 estimating area and assessing accuracy of land change, *Remote Sens. Environ.*, 148, 42–57,
1147 <https://doi.org/10.1016/j.rse.2014.02.015>, 2014.
- 1148 Pan, L., Xia, H., Yang, J., Niu, W., Wang, R., Song, H., Guo, Y., and Qin, Y.: Mapping cropping intensity in Huaihe
1149 basin using phenology algorithm, all Sentinel-2 and Landsat images in Google Earth Engine, *Int. J. Appl. Earth*
1150 *Obs. Geoinf.*, 102, 102376, <https://doi.org/10.1016/j.jag.2021.102376>, 2021.
- 1151 Ren, J., Shao, Y., and Wang, Y.: Annual Cropping Intensity Dynamics in China from 2001 to 2023, *Remote Sens.*,
1152 16, 4801, <https://doi.org/10.3390/rs16244801>, 2024.
- 1153 Waha, K., Dietrich, J. P., Portmann, F. T., Siebert, S., Thornton, P. K., Bondeau, A., and Herrero, M.: Multiple
1154 cropping systems of the world and the potential for increasing cropping intensity, *Glob. Environ. Change*, 64,
1155 102131, <https://doi.org/10.1016/j.gloenvcha.2020.102131>, 2020.
- 1156 Yang, G., Yu, W., Yao, X., Zheng, H., Cao, Q., Zhu, Y., Cao, W., and Cheng, T.: AGTOC: A novel approach to
1157 winter wheat mapping by automatic generation of training samples and one-class classification on Google Earth
1158 Engine, *Int. J. Appl. Earth Obs. Geoinf.*, 102, 102446, <https://doi.org/10.1016/j.jag.2021.102446>, 2021.
- 1159 You, N., Dong, J., Li, J., Huang, J., and Jin, Z.: Rapid early-season maize mapping without crop labels, *Remote*
1160 *Sens. Environ.*, 290, 113496, <https://doi.org/10.1016/j.rse.2023.113496>, 2023.
- 1161 Zhang, H., Song, M., Ding, Y., Xie, Y., Guo, H., Xu, L., Ge, J., Zhu, Y., Wang, S., and Guo, Z.: GlobalRice20: A
1162 20 m resolution global paddy rice dataset for 2015 and 2024 derived from multi-source remote sensing, *Earth*
1163 *Syst. Sci. Data Discuss.* [preprint], <https://doi.org/10.5194/essd-2026-24>, 2026.
- 1164 Zhang, M., Wu, B., Zeng, H., He, G., Liu, C., Tao, S., Zhang, Q., Nabil, M., Tian, F., Bofana, J., Beyene, A. N.,
1165 Elnashar, A., Yan, N., Wang, Z., and Liu, Y.: GCI30: a global dataset of 30 m cropping intensity using
1166 multisource remote sensing imagery, *Earth Syst. Sci. Data*, 13, 4799–4817, <https://doi.org/10.5194/essd-13-4799-2021>, 2021.