Author's Thanks:

We sincerely appreciate the reviewer's dedicated time and expertise in critically evaluating our work. The constructive feedback has prompted essential refinements to both the scholarly substance and structural clarity of this manuscript, significantly elevating its academic contribution. Below we provide a systematic point-by-point response to each comment. The italicized content represents the modifications made in the manuscript.

RC1 Comment 1:

Field data is critical for model training and map validation. However, the manuscript lacks clarity and details regarding how the authors collected the field samples and how they chose the field target for visual interpretation through GEE. The sampling strategy for field data collection is unknown, especially for validation data collection. Using points to validate 30-m maps is inappropriate especially when mixed pixels occur. The reported accuracies could be largely impacted by 30-m mixed pixels when only point data are employed for map validation.

Author's Response 1:

Thank you for your valuable suggestion. To ensure comprehensive representation of the validation dataset, the study area was stratified into major paddy cultivation areas and non-paddy regions. Ground truth data were randomly sampled from each stratum, maintaining an approximate 3:2 ratio between paddy and non-paddy areas. This resulted in a total of 68856 paddy samples and 39098 non-paddy samples, as illustrated in Fig.1(b). Specifically, the dataset includes 21254 paddy samples and 13160 non-paddy samples obtained from field survey, complemented by 47602 paddy samples and 25938 non-paddy samples from Google Earth VHR imagery. To mitigate the effects of mixed pixels in the validation results, detailed field observation data was conducted within a 30m radius around each sample point, recording the proportions of coverage for paddy rice. And the confusion matrix of was calculated in terms of the proportion of area. This comprehensive dataset was used to evaluate the accuracy and reliability of the paddy rice mapping results.

RC1 Comment 2:

For the map evaluation, the distribution of the validation dataset is not reported in the manuscript. Are they derived from probability sample? Otherwise, the validation would not be valid. Constructing a confusion matrix based on pixel counting is not recommended. The population error matrix of classes with cell entries should be expressed in terms of the proportion of area. Besides,

the uncertainty of these accuracy metrics should also be reported. Refer to Olofsson et. al (2014) for a guideline on how to conduct map accuracy evaluation solidly.

Author's Response 2:

Thanks for your valuable comment. The validation dataset was developed through a stratified random sampling approach. To reduce the influence of mixed pixels (paddy and other types) and according to the methodology proposed by Olofsson et al., (2014) we refined the confusion matrices presented in Tables 2 and 3 to reflect area-based proportions. Furthermore, detailed descriptions of the evaluation metrics and their corresponding formulas have been incorporated into Section 2.5, while the report of accuracy indicator uncertainties has been included in Section 4.3. Besides, Tables 4, 5, and 6 in the Discussion section evaluates the accuracy of FR-Net models based on training set, no ground truth data; therefore, no changes are needed.

Reference:

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

RC1 Comment 3:

L22: Are there average numbers from 1985 to 2023? Make it clearer.

Author's Response 3:

Thanks for your valuable comment. These values represent the average accuracy evaluation indicators for paddy rice maps from 1985 to 2023, and we have clarified these in the manuscript. The modified content is as follows: "*The overall mapping result obtained from the FR-Net model and ARE methods achieved high average values of user accuracy (UA) of paddy, producer accuracy (PA) of paddy, overall accuracy (OA), F1 score, and Matthews correlation coefficient (MCC) values of 0.93, 0.91, 0.91, 0.92, and 0.82, respectively.*"

RC1 Comment 4:

L44-45: What is the justification for such a statement that it's challenging to produce long-term maps using phenology-based methods? any references?

Author's Response 4:

Thanks for your valuable comment, we have modified the corresponding content in the manuscript. The modified content is as follows: "*While this method is simple and effective, cloud cover can*

cause gaps in satellite data during the growth period, increasing the difficulty to track phenological changes accurately. This limitation can reduce the accuracy of long-term rice mapping. (Carrasco et al., 2022; Dong et al., 2015)."

Reference:

Carrasco, L., Fujita, G., Kito, K., Miyashita, T., 2022. Historical mapping of rice fields in Japan using phenology and temporally aggregated Landsat images in Google Earth Engine. ISPRS J. Photogramm. Remote Sens. 191, 277–289.

Dong, J., Xiao, X., Kou, W., Qin, Y., Zhang, G., Li, L., Jin, C., Zhou, Y., Wang, J., Biradar, C., Liu, J., Moore, B., 2015. Tracking the dynamics of paddy rice planting area in 1986–2010 through time series Landsat images and phenology-based algorithms. Remote Sens. Environ. 160, 99–113.

RC1 Comment 5:

L64: I think you are saying that the final map for a specific year is derived from multiple intermediate maps within the year. But 'multiple annual results" means multiple yearly maps, i.e., a map for each year. This could be confusing.

Author's Response 5:

Thank you very much. We believe 'intermediate maps' is a better option. According to your guidance, we have revised it in the manuscript. As below: "However, determining the final mapping result from multiple annual results remains a challenge for large-scale paddy rice mapping." to "*However, determining the final mapping result for a specific year from multiple intermediate maps remains a challenge for large-scale paddy rice mapping.*"

RC1 Comment 6:

L116: You mentioned Result_pre in the text but there is no Result_pre in Eq.1.

Author's Response 6:

Thanks for your valuable comment. We are sorry for the careless, and the Eq. (1) has been revised accordingly.

"$Result_{pre} = \begin{cases} nonpaddy. else \\ paddy, \forall i \in [1, \dots m]: Result_i = paddy \end{cases}$, (1)"

RC1 Comment 7:

L132: From Eq.2, t represents the image corresponding to the highest absolute value of the difference between the category probability and 0.5, not the direct highest Pi. Why not use max (Pi) instead? For example, if P1=0.1, P2=0.6, then there would be t=1, and Pt=P1=0.1. Would you

determine the final results as non-paddy since Pt < 0.5?

Author's Response 7:

Thanks for your valuable comment. If P1=0.1, P2=0.6, we would determine the final results as non-paddy. This is because we set 0.5 as the classification threshold. |P1-0.5|=0.4 indicates a higher confidence level in classifying the sample as non-paddy, whereas |P2-0.5|=0.1 suggests a lower confidence level in identifying it as paddy, accompanied by higher classification uncertainty. To enhance the robustness of the classification result, we prioritize the feature with stronger confidence. Consequently, the final category is determined as non-paddy in this scenario. If max (P_i) is used, misclassification issues may occur, such as incorrectly identifying non-paddy as paddy. For example, if P1=0.1, P2=0.2, and P3=0.6, the category probability of being non-paddy is higher. In this case, using max(P_i) to classify the final result as paddy would be erroneous.

RC1 Comment 8:

L136: Are you using exactly the same parameters (models) in these different phenological stages? Otherwise, how could you ensure that the category probability outputs among m images are comparable?

Author's Response 8:

Thanks for your valuable comment. We share the same point. Exactly the same parameters (models) were used across all phenological stages. Only in this way can we ensure the category probability outputs of the m images are comparable. We supplemented it in the manuscript: "*In this study, all FR-Net models utilized the same set of parameters for subsequent analysis.*" Thank you.

RC1 Comment 9:

L142: In which months did you download the data and from where? Specify the date range for each year, or the same range across years if they are consistent.

Author's Response 9:

Thanks for the insightful comment. We obtained Landsat Collection 2 Level-2 surface reflectance products for Northeast China from the United States Geological Survey (USGS), covering the period from May to September annually between 1985 and 2023. A total of 13,809 images were downloaded, each with a spatial resolution of 30 meters. The aforementioned content has been incorporated into the manuscript.

RC1 Comment 10:

L145: Please provide the specific band names instead of numeric names.

Author's Response 10:

Thanks for your valuable comment, we have modified the corresponding content in the manuscript. "*In addition, we selected the Blue, Green, Red, Near Infrared (NIR), Shortwave Infrared (SWIR) 1, and Shortwave Infrared (SWIR) 2 bands of Landsat 8/9 OLI and Blue, Green, Red, NIR, SWIR 1, and SWIR 2 bands of Landsat 5 TM images to map the paddy rice.*"

RC1 Comment 11:

L156: There are some issues with the validation dataset. The sampling strategy is unknown. How did you select the field sites to visit? If the distribution of validation datasets is biased (not randomly selected), then the map accuracy based on the validation datasets is not valid. Did you collect field data as point observations? Using points to validate 30-m maps is inappropriate especially when mixed pixels occur. What are the spatial and temporal distributions of training and validation data?

Author's Response 11:

Thanks for your valuable comment. To ensure the accuracy of the validation dataset, field samples were collected in major paddy cultivation areas and non-paddy regions. The spatial distribution of these validation samples is depicted in Figure 1(b), with stratified random sampling applied to randomly select samples within each stratum. A total of 34414 field survey data and 73540 VHR data were collected as point observations. The sampling time of field and VHR data aligned with the Landsat image acquisition period (May to September) to ensure temporal consistency. To mitigate mixed pixels effects in validation results, detailed field observations were conducted within a 30m radius around each sample point, recording the proportion coverage of paddy rice. And the confusion matrix of was calculated in terms of the proportion of area.

RC1 Comment 12:

L170: 29906 + 9968 + 50956 + 16985 = 107815, this is less than the total size (68865 + 39098 = 107963, L154), did you remove any ground samples and why? What are the criteria for dividing the entire ground data into these training/validation sets with these specific numbers, for Landsat5 and 8/9 respectively?

Author's Response 12:

Thanks for your valuable comment. In this study, no ground samples were removed. The validation data (ground truth) comprised 107963 field survey and Very High Resolution (VHR) samples, while

107815 Landsat images, each with a size of 256×256 pixels, were used to train the model.

RC1 Comment 13:

L205: Please use explicit band names. In Fig.3, did you use the same probability threshold of 0.5 in both overlay maps and the ARE maps? For the red circled area, e.g., in E5, a non-paddy pixel in the overlay map means all category probability outputs are less than 0.5. According to Eq2, for ARE methods, a paddy pixel must have a probability greater than 0.5. How come a non-paddy pixel in the overlay map would become a paddy pixel in the ARE map?

Author's Response 13:

Thanks for your valuable comment, we have modified explicit band names in Lines 205-206. In Fig.3, we used the same probability threshold of 0.5 in both overlay maps and the ARE maps. Regarding E5 in Fig.3, due to manuscript length constraints and to emphasize the differences between the overlay results and ARE results, only the results from columns A and C were displayed. However, the results in column F were generated using all available images from May to September for that year. To further clarify, we have added column G, which depicts the ARE results calculated from the results in columns A and C.

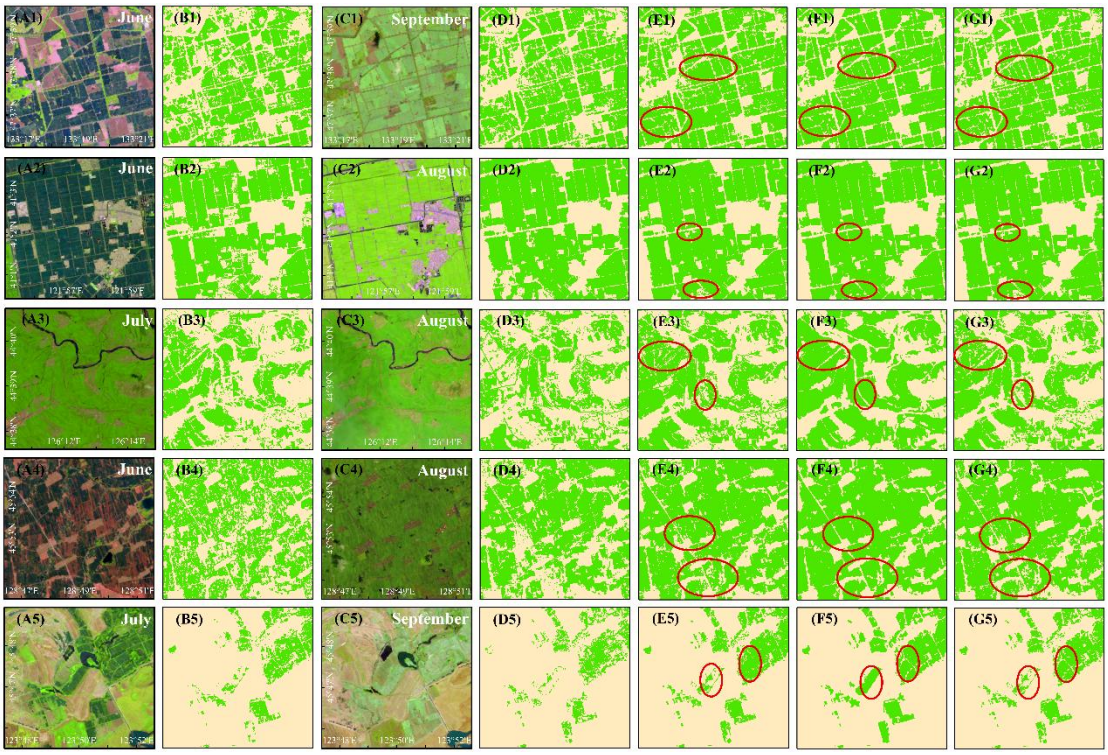

Figure 3: Comparison of paddy rice maps between ARE, single temporal, and overlay methods using Landsat 5 TM and Landsat 8 OLI images. (A1)-(A3), (C1)-(C3) represent pseudo-colored maps of three regions in the Landsat 8 images (bands SWIR 1, NIR, and Red) from June to September. (A4), (A5), (C4), and (C5) represent pseudo-colored maps of two regions in the Landsat 5 images (bands SWIR 1, NIR, and Red) from June to

September. B_i display the paddy rice maps corresponding to the images of A_i. D_i shows the paddy rice maps corresponding to the images of C_i. E_i are the overlay results of the paddy rice maps from A_i and C_i. F_i depict ARE paddy rice maps obtained from all images available at that year. G_i depict the ARE paddy rice maps from A_i and C_i, $i \in [1,5] \cap \mathbb{Z}$. The red circles in the overlay and ARE maps indicate areas with significant differences in paddy rice.

RC1 Comment 14:

L212: The distribution of the validation points is totally unknown. Are they derived from probability samples? Otherwise, the validation would not be valid. A confusion matrix based on pixel counting is not recommended. The population error matrix of classes with cell entries should be expressed in terms of the proportion of area. Besides, the uncertainty of these accuracy metrics should also be reported. Refer to Olofsson et. al (2014) for a guideline on how to conduct map accuracy evaluation in a solid manner.

Reference:

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

Author's Response 14:

Thanks for your valuable comment. We employed a stratified random sampling strategy to select ground validation samples across major paddy cultivation areas and non-paddy areas within the study area, with a total of 107954 samples. Furthermore, to account for the influence of mixed pixels, we reformulated the confusion matrix based on the proportion of area according to the study of Olofsson (Olofsson et al., 2014), which has been meticulously revised in the manuscript.

Reference:

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

RC1 Comment 15:

L236: Fig.4, what is the scale of this comparison? I assume this is the total area in the entire study area. What about the comparisons at the district, municipal, and provincial levels since you collected the agricultural statistics?

Author's Response 15:

Thanks for your valuable comment. In Fig.4, we validated the total area of paddy rice from 1985 to 2023 using agricultural statistical data across the entire study area. Owing to the unavailability of statistical data for certain counties and municipalities, we utilized all publicly available agricultural statistical data from the study area to enhance the validation of paddy rice mapping results at the provincial, municipal, and district levels in the manuscript.

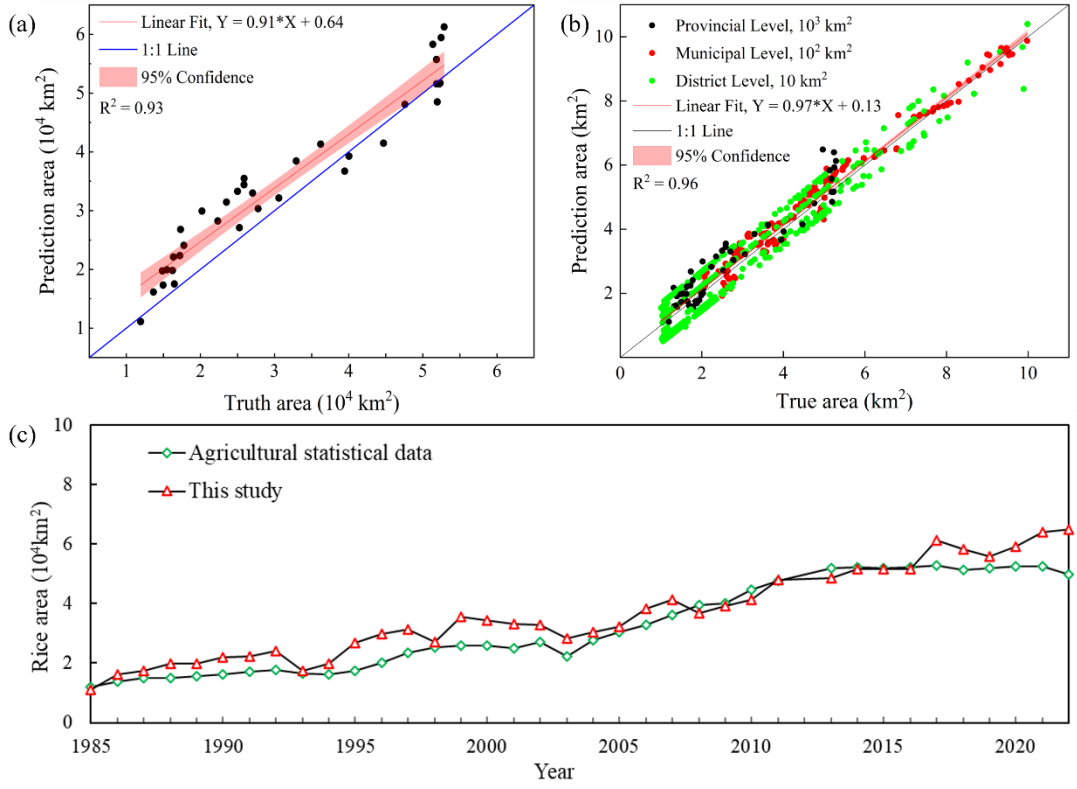

Figure 4: Comparison of paddy rice maps with agricultural statistical data. (a) presents the paddy rice area comparison across the entire study area. (b) is the comparison of the paddy rice area at provincial, municipal, and district levels within the study area. (c) shows the difference between the agricultural statistical data and the results of this study regarding paddy rice area estimation.

RC1 Comment 16:

L263: what if there are no clear-sky observations available in the proceeding and subsequent years? did you leave it as no data? In Fig.7, what is "interpolated paddy"? Did you interpolate your classification directly from the classification in the previous/next year, if there is no cloud-free satellite data in the current year? This has to be clarified.

I assume the interpolated paddy pixels are derived from satellite data that are interpolated from previous/next Landsat observations. The term 'interpolated paddy' implies that the classified pixel itself is somehow interpolated, which makes no sense.

Author's Response 16:

Thank you for your valuable comment, we have modified the corresponding content in the manuscript. After analyzing and processing the Landsat images from 1985 to 2023 in Northeast China, no sustained deficiency in clear-sky observation data availability has been identified within the study area over the examined period. If the method proposed in this study is applied to other regions facing this issue, incorporating multi-source data such as MODIS, Sentinel-1, and Sentinel-2 could be considered to achieve a more comprehensive analysis. And the term 'interpolated paddy' refers to the paddy rice result derived using a multi-year comprehensive method, based on the historical phenological patterns from the nearest available clear-sky year's image. In addition, to avoid ambiguity, we replaced 'interpolated paddy' with 'gap-filled paddy' throughout the manuscript.

RC1 Comment 17:

L275-276: There is no figure showing the admin boundaries and labels. It's hard to reader unfamiliar with China to link your descriptive context to the spatial locations in the map.

Author's Response 17:

Thanks for your valuable comment, we have modified Figure 8 in the manuscript.

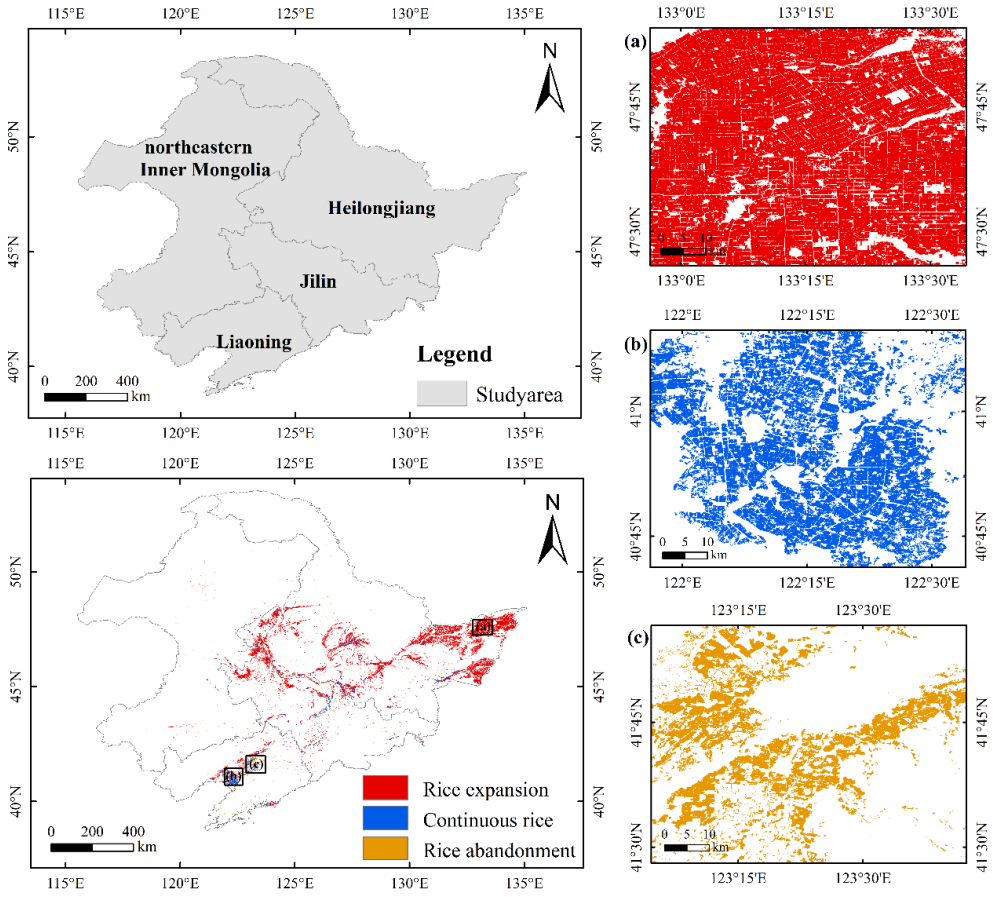

Figure 8: Trends of paddy rice cultivation area in Northeast China from 1985 to 2023.

RC1 Comment 18:

L301: This is contradictory to the table. Instead, #1 and #4 show that using data from only one sensor achieved the best accuracy and had the best results (if the models are trained and applied to the same sensor), compared to other scenarios using multiple sensors.

Author's Response 18:

Thanks for your valuable comment, we have revised the corresponding content in the manuscript. The modified content follows: "*This means models trained using single-sensor images cannot be effectively applied to images from other sensors (combination numbers 2 and 3). Therefore, to ensure reliable large-scale mapping and longitudinal consistency, we advocate for systematic cross-sensor fusion strategies rather than single-sensor dependencies, thereby mitigating the differences in feature distributions between sensors.*"

RC1 Comment 19:

L304-305: Not clear how you conducted 'transfer learning". For #8, training on L5 & 20% L8, then apply to L8? Need to clarify.

Author's Response 19:

Thanks for your valuable comment. Sorry for the misleading, and we have clarified the issue you raised in the manuscript.

For #8, the model trained with combination number 1 was further trained using 20% of Landsat 8/9 OLI images to achieve subtle adjustments in the model weights, which were then applied to the remaining Landsat 8/9 OLI images.

RC1 Comment 20:

L306: An enhanced accuracy compared to what?

Author's Response 20:

Thanks for your valuable comment, we have modified the corresponding content in the manuscript. "The results revealed that fine-tuning the models enhanced accuracy compared to combination numbers 2 and 3."