Dear editor,

Thank you for your comments concerning our manuscript (MS). We have substantially revised our manuscript with the comments provided by the reviewer. We have studied comments carefully and have made corrections, which we hope meet with approval. Revised portions are marked in red on the paper. The leading corrections in the paper and the response to the reviewer's comments are as follows:

Generating spatial crop yield information is of great significance for academic research andguiding agricultural policy. Here Zhao et al., generated a 30-m winter wheat yield dataset covering main winter wheat-growing region of China from 2016 to 2021. The authors said they proposed a a semi-mechanistic model named HLM that showed better performance against one commonly used machine learning named random forest (RF). With the model developed, and observational dataset including meteorological variables, vegetation index, and yield, they generated the grid-level yield dataset across multiple years. I think this work is important and I would recommend it for publication if my following major concerns, mainly related to model configuration, evaluation, and comparison, could be resolved. The detailed comments are listed as bellows.

Major comments

1) Clarify the model used: section 2.3.1, Eq. 2, what does j represent? Did you develop each HLM separately for each province or city over each month? In Eq.3, what does βmj represent in Eq. 2? Did you normalize each input variable before inputting those variables to the model? How did you solve the parameters in Eq.2-3? Please clarify those details and code to make those results reproducible.

[Response]: Thank you very much for your suggestion. The hierarchical linear model (HLM) is a simple and efficient method for dealing with nested structures. In this study, normalization was performed on the data before modeling to reduce the impact of differences in variable scales. For each province, a set of parameters was generated by using the data collected from the sample fields. The specific yield-predicting models in different provinces using the HLM method in this study involved a two-levels hierarchy. βj represents the β0 and β1 from Level 1 of HLM, j represents 0 or 1. The parameters of the HLM model in this article are estimated using maximum likelihood estimation. Modification is incorporated in Line 201-206.

2) Baseline machine learning model configuration: Line 216-219: did you build one RF model for the entire study area and compare it with multiple HLM models? If so, it should not be a fair comparison.

What if you built multiple RF models? In addition, how did you select the key parameters used in RF model, including maximum depth of the tree, the number of features, minimum number of samples required to split an internal node, and minimum number of samples required to be at a leaf node. Note that inappropriate model configuration would generate a worse performance for the RF model. Please clarify how did you select those parameters in details.

[Response]: Thank you for asking, we generated multiple RF models for each province just like the way we build HLM models, using same calibration and validation datasets, so it makes two models for each province and definitely comparable. To clearly demonstrate our approach, the corresponding L224-232. Concerning the key RF parameters, we optimized the models' hyperparameters through pretuned procedure, using 10-fold cross-validation. Majorly we adjust the number of trees and was tuned to 200 trees. There are several relevant studies have indicated the similar results(Li et al., 2021; Cheng et al., 2022). In the data analysis of this study, it has been found that the RF model achieves stable accuracy and small errors with a number of trees below 200.

Cheng, M., Jiao, X., Shi, L., Penuelas, J., Kumar, L., Nie, C., Wu, T., Liu, K., Wu, W., and Jin, X.: High-resolution crop yield and water productivity dataset generated using random forest and remote sensing, Scientific Data, 9, 641, 10.1038/s41597-022-01761-0, 2022.

Li, L., Wang, B., Feng, P., Wang, H., He, Q., Wang, Y., Li Liu, D., Li, Y., He, J., and Feng, H.: Crop yield forecasting and associated optimum lead time analysis based on multi-source environmental data across China, Agricultural and Forest Meteorology, 308, 108558, 2021.

3) Model validation: since the authors said that the proposed models can overcome inter-annual and cross-regional problems. It's therefore required to validate the model performance by cross-region and cross-year. Line 235-237, did you train the model over part of the studied regions and then validate model performance over the remaining part of the studied regions where the model has never seen the observed yield values (i.e., cross-regional validation to show model overcome the cross-regional problem)? In addition, experiments are required to validate the model performance by training the model in specific years and then validating the model performance in other years that the model has never seen its observed yield data (i.e., cross-year validation to show model overcome the inter-annual problem).

[Response]: Thank you very much for your suggestion. The cross-validation results are based on modeling in a certain year and verification in other years. Region cross-validation models one region and validates other regions. In this study, regional and temporal cross-validation was performed by training

the models on specific years or regions and then independently validating them on the remaining years or study regions as separate samples. Modification is incorporated in Line 253-255, Line 296-298, Fig.7 and Fig.8.

4) Fig.5-6, did you include all the training and validation dataset in this scattered plot or only used the validation dataset for this evaluation? Please clarify it in the main text or in the figure caption. Note that the model performance should be validated against independent validation dataset that the model has never seen. In Fig. 5, please clarify what does each point represent.

[Response]: Thank you very much for your suggestion. In this paper, the data were randomly split into two dataset, two-thirds of the data were used for modelling, and the remaining data were used for validation. This article uses independent samples for model validation, which means that the data used for validation are not included in the modeling data. The titles of Figure 5 and Figure 6 have been updated. Figure 5 presents the results based on the modeling dataset samples, while Figure 6 shows the results based on independent validation data. Additionally, the different colored dots in Figure 5 have also been explained.

5) Fig.5-Fig.7, why not compare RF and HLM using the validation schemes in Fig.5-7? If no such kind of comprehensive comparison, it could be less convincing to draw the conclusion that the proposed method had excellent accuracy.

[Response]: Thank you very much for your suggestion. For the reliability of model comparison, RF and HLM were compared using the same modeling dataset and validation dataset. The results showed that RF had higher accuracy on the modeling dataset, but its stability on the validation results was poor. Therefore, HLM was used for subsequent result presentation and generation of yield datasets. In order to reduce reader misunderstanding, we have reorganized the data analysis results. Modification is incorporated in Fig.3, Fig.5, Fig.6 and Line 281-289.

6) In addition to '**', please explicitly show the p-value for all related results.

[Response]: Thank you very much for your suggestion. In this article, ** represents model significance at the 0.01probability level ($p < 0.01$). The relevant figures, tables, and descriptions in the article have all been renumbered.

Specific comments

1) Line 19, "a coarse spatial resolution" ranging from what? Ranging from 100 meters to 1 km? Since

the major uniqueness of the dataset is its high spatial resolution, I suggested to show the spatial resolution of previous related dataset in the abstract with few words.

[Response]: Thank you very much for your suggestion. Indeed, showcasing the resolution of existing yield datasets helps highlight the high spatial resolution of this study. Based on your feedback, we have incorporated relevant information into the abstract. Modification is incorporated in Line 19-20.

2) Line 19-21, the sentence is less rigorous. I would revise it as "useful for analyzing large-scale temporal and spatial changes in yield, …deal with small-scale spatial heterogeneity, which …of the Chinese farmers' economy"

[Response]: In accordance with your suggestions, the sentence can be modified as follows: "Although these datasets are useful for analyzing large-scale temporal and spatial change in yield and they cannot deal with small-scale spatial heterogeneity, which happens to be the most significant characteristic of the Chinese farmers' economy." Thank you very much for your suggestion. Modification is incorporated in Line 19-22.

3) Line 29-31, I understand that '**' could be related to significance test (i.e., p-value), but I suggest to remove those symbols in the abstract since you even did not explain it. Note that the same symbol with no explanation could represent different meanings which could confuse the readers.

[Response]: Thank you very much for your suggestion. Based on your feedback, all the "**" symbols have been removed as suggested.

4) Line 68-69, "ML methods need large multidimensional datasets, which can challenge their application", that's not the limitation of ML methods. Note that the inputs of ML models can be the same to parametric regression models or the process-based models. Additionally, there have been many strategies developed for dealing with multi or high-dimensional inputs in ML.

[Response]: Thank you very much for your suggestion. Yield estimation requires a large amount of data, but data generation algorithms have already been developed in ML methods. Based on your feedback, the sentence can be modified as follows: "Overall, ML methods heavily rely on large training datasets (Cao et al., 2021). Nonetheless, the application of machine learning in the realm of synthetic data generation has also exhibited encouraging outcomes (Arslan et al., 2019; Sivakumar et al., 2022; Ebrahimy et al., 2023)". Modification is incorporated in Line 69-71.

5) Line 91, revise "migration learning" as "transfer learning" that is commonly named as

[Response]: Thank you very much for your suggestion. We have revised “migration learning” as “transfer learning”. Modification is incorporated in Line 93.

6) Line 99, revise “county to city scale” as “county or city scale” since a county can be larger than a city in US.

[Response]: Thank you very much for your suggestion. We have revised “county to city scale” as “county or city scale”. Modification is incorporated in Line 101.

7) Line 109-111, please see my comment#2

[Response]: In accordance with your suggestions, the sentence can be modified as follows: “Although these datasets are useful for analyzing large-scale temporal and spatial change in yield and they cannot deal with small-scale spatial heterogeneity, which happens to be the most significant characteristic of the Chinese farmers' economy.” Thank you very much for your suggestion. Modification is incorporated in Line 111-113.

8) Fig. 1, remove the explanation of “sampling points” in the title of the figure since it has been explained in the figure legend. In addition, clear reasons need to be given for selecting those three regions. E.g., why they represent crop yield at different kinds of background conditions? How do you define those background or climate conditions?

[Response]: Thank you very much for your suggestion. (1) We remove the explanation of “sampling points” in the title of the figure 1. The three selected regions in this study were chosen for comparison with other yield datasets based on different wheatland coverages. Region 1, 2, and 3 represent areas with winter wheat coverages below 25%, around 50%, and above 75%, respectively, serving as representative regions for these respective coverage levels. Modification is incorporated in Fig.1 and Line 136-139.

9) Section 2.2.2, top of atmosphere or surface reflectance was used? Any data quality controls were applied? Please clarify.

[Response]: The GEE platform stores atmospherically corrected reflectance data from Sentinel 2 and Landsat 8, which is the dataset we used (Figure 3). In order to provide a clearer understanding of our dataset to the readers, corresponding modifications have been made in Line 152-155, as follows: “In this work, we extracted the atmospherically corrected reflectance from Landsat 8 and Sentinel 2 images on the Google Earth Engine (GEE) platform during the period of 2016-2021. Subsequently, we calculated the Enhanced Vegetation Index 2 (EVI2) (Jiang et al., 2008) using the extracted reflectance values”.

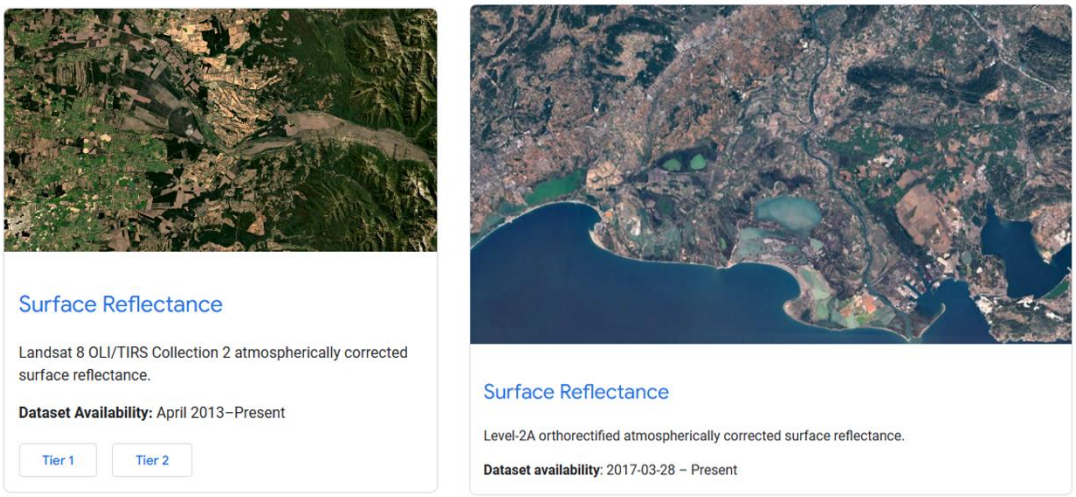

Figure 3 Reflectance dataset from Landsat8 and Sentinel2 based on the GEE platform.

10) Section 2.2.3, is there any high-resolution meteorological data in China? why not use those high-resolution data rather than ERA5?

[Response]: Thanks for your suggestions. (1) ECMWF atmospheric reanalysis refers to data assimilation of near-surface observation data to obtain raster data, which will cover a larger area. This reanalysis combines observations into globally consistent fields taking into account the dynamics and physics of the model using a data assimilation process (four-dimensional variational analysis, 4D-Var, in the case of ERA5). (2) Figure 1 shows the distribution of weather stations in the Huang-Huai-Hai region. It is clear that the existing weather data cannot cover every county. In the process of data analysis, we compared the difference between the weather station data and ECMWF weather data (Figure 2). The results showed that the ECMWF data has a high consistency with the measured data. Similar phenomena occur in other regions. Therefore, ECMWF is a reasonable analysis data set. (3) In our previous research, we have also demonstrated the stability of applying ECMWF data in vegetation monitoring and yield prediction (Xu et al., 2020; Li et al., 2020; Zhao et al., 2022). (4) ERA5 data can be accessed on the GEE platform, which provides convenience for subsequent regional-scale yield prediction. Based on the above, we have chosen to use ERA5 data for subsequent analysis.

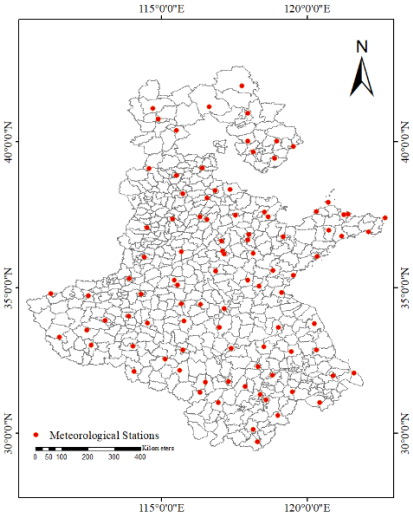

Figure 1 Distribution of Meteorological Stations in Huang-Huai-Hai region

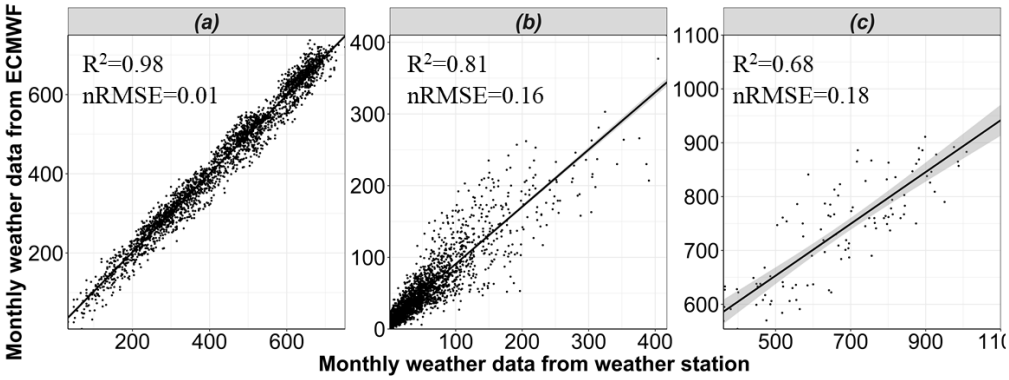

Figure 2 Comparison of ECMWF weather data with measured data from weather stations: (a) monthly accumulated temperature (℃), (b) monthly accumulated precipitation (mm) and (c) monthly accumulated radiation (MJ m-2).

Xu, X., Teng, C., Zhao, Y., Du, Y., Zhao, C., Yang, G., Jin, X., Song, X., Gu, X., Casa, R., Chen, L., Li, Z.: Prediction of wheat grain protein by coupling multisource remote sensing imagery and ECMWF data. Remote Sensing, 2020, 12: 1349.

Li, Z., Taylor, J., Yang, H., Casa, R., Jin, X., Li, Z., Song, X., Yang, G. A hierarchical interannual wheat yield and grain protein prediction model using spectral vegetative indices and meteorological data. Field Crop Res 2020, 248, 107711, https://doi.org/10.1016/j.fcr.2019.107711.

Zhao, Y., Han, S., Meng, Y., Feng, H., Li, Z., Chen, J., Song, X., Zhu, Y., Yang, G. Transfer-Learning-Based Approach for Yield Prediction of Winter Wheat from Planet Data and SAFY Model. Remote Sens. 2022, 14, 5474. https://doi.org/ 10.3390/rs14215474

11) Line 226-227, how did you calculate the average yield using ChinaWheatYiled30m? was it calculated by first summarizing the yield values of pixels within the study area, and then divide by the number of pixels?

[Response]: Thanks for your suggestions. The provincial and municipal average yields based on the ChinaWheatYield30m dataset were calculated by dividing the total yield of all winter wheat pixels by the number of winter wheat pixels in that area. Modification is incorporated in Line 241-243.

12) Fig. 3, what does each point represent? Please clarify it in the figure caption

[Response]: Thank you very much for your suggestion. In order to better display the results, Figure 3 has been changed to a histogram. Modification is incorporated in Fig.3.