Responses to the comments of Referee #2

Article ID: essd-2024-586
Title: NortheastChinaSoybeanYield20m: an annual soybean yield dataset at 20 m in Northeast China from 2019 to 2023
Authors: Jingyuan Xu, Xin Du, Taifeng Dong, Qiangzi Li, Yuan Zhang, Hongyan Wang, Jing Xiao, Jiashu Zhang, Yunqi Shen, Yong Dong

Dear Reviewer,

Thank you very much for your thorough review and constructive feedback on our manuscript. We have carefully addressed each comment and suggestion to refine our work, enhance its clarity and strengthen its scientific contribution. The key revisions include:

(1) Strengthened the Introduction section. We have restructured the introduction to better contextualize the critical issues on coupling data-driven and knowledge-drive methods for crop yield estimation. The revised edits now explicitly highlight the limitations of the existing models (e.g., coarse resolution, and over-reliance on ground data).
(2) Expanded details in data processing. We expanded the Data collection section to include details on data processing procedures especially for the meteorological and satellite imagery data.
(3) Enhanced interpretation of results. We strengthened the Results and Discussion sections by analyzing yield estimation uncertainty across different scales and discussing key sources of error, providing deeper insights into model performance and study implications.
(4) Quantified MODIS-Sentinel-2 Comparison in yield estimation. We have added a new subsection in the Discussion section on quantitative comparison of the performance of MODIS LAI and Sentinel-2 data in yield estimation.

The detailed point-to-point responses are as follows. Texts in black are the reviewer's comments; those in blue are our responses to the reviewer's comments; and those in *red and italics* are the revised texts appeared in the revised manuscript.

The overall structure of the article is clear and logically organized. The research demonstrates innovation by integrating crop growth models with deep learning algorithms for soybean yield estimation, representing a promising direction in agricultural remote sensing. The research objectives are well-defined, aiming to address existing limitations in soybean yield data (insufficient spatial resolution and reliance on ground observations), thereby supporting optimized soybean production distribution and agricultural decision-making.

Reply: Thank you for your positive feedback and recognition of our work. In the revision, we have carefully addressed your thoughtful comments and suggestions to improve our manuscript.

Specific Comments:

1 Introduction: The section comprehensively highlights soybean's global food security significance and limitations of current yield estimation methods, establishing a solid research rationale. However, the comparative discussion of data-driven and knowledge-driven methods could be more concise to better emphasize core issues and proposed solutions. Additionally, enhancing explanations of environmental factors' mechanisms (e.g., how climatic conditions affect growth cycles and photosynthesis, or how soil properties constrain nutrient uptake and water retention) would provide a more systematic understanding of key yield determinants and their interactions.

Reply: Thank you for your valuable comments.

(1) In the revised revision, we have refined the statements on the advantages and limitations of existing methods, placing greater emphasis on our proposed method.
(2) We discussed the impact of environmental factors (climate conditions and soil properties) on crop growth in the introduction. Specifically, we clarified the limitations of data-driven methods in accounting for environmental factors, and highlighted the strengths of knowledge-driven models incorporating these influences.

Revisions can be found in Section 1 Introduction. Below is a part of the revision for your reference:

Data-driven methods leverage satellite-derived variables such as leaf area index (LAI), fraction of absorbed photosynthetically active radiation (FAPAR), and vegetation indices (VIs) to establish linear or nonlinear relationships with measured crop yield (Ang et al., 2022; Xie et al., 2019). Machine learning algorithms such as Random Forest (RF), and Artificial Neural Networks (ANN), due to their ability to process large dataset and model complex nonlinear interactions, have been widely applied in crop yield estimations (Pang et al., 2022; Tian et al., 2021; Yildirim et al., 2022). These methods can extract effective information from multi-source structured

or unstructured data without manual intervention. However, they are heavily reliant on extensive ground-truth training data, which is challenging to collect over large areas and high time intervals (Cao et al., 2021). ***Additionally, these models often overlook the impacts of environmental factors on crop growth, such as the influence of early-season soil moisture on root establishment or the effect of high temperatures during flowering on pod set, and are lack of interpretability, as they cannot explain the causal relationship between input features and outputs, leading to poor spatial-temporal generalization (Gevaert, 2022).***

In contrast, knowledge-driven crop growth models simulate crop development from sowing to harvest based on agronomic mechanisms (Kaur and Singh, 2020). Common model types include light-use efficiency models (e.g., SAFY (Duchemin et al., 2008)), soil-driven models (e.g., AquaCrop (Steduto et al., 2009)), and atmospheric-driven models (e.g., WOFOST (Diepen et al., 1989)). ***These models integrate environmental factors (e.g., climate conditions and soil characteristics) with crop physiological processes (Gaso et al., 2024). Climate variables like temperature, precipitation, and solar radiation are critical in regulating essential physiological processes such as photosynthesis, respiration and transpiration, which influence the rate and duration of crop growth stages (Misaal et al., 2023). Climate anomalies during specific growth stages may disrupt biochemical processes, ultimately affecting yield formation. Similarly, soil properties influence crop productivity by regulating water retention, aeration, and nutrient uptake (Muhuri et al., 2023).*** *Despite their mechanistic rigor, applications of crop models over large area are typically constrained by (1) insufficient spatial-temporal input data, and (2) parameter uncertainty, which can propagate errors into yield estimations (Dokoohaki et al., 2021). To overcome these challenges, data assimilation techniques to integrate remote sensing observations (e.g., LAI) into crop growth models have been developed to enhance spatial representativity (Huang et al., 2024). However, high resolution remote sensing data drastically increases computational cost, limiting the scalability of these approaches for regional or national mappings efforts (Huang et al., 2019).*

Given the limitations above, integrating data-driven and knowledge-driven models has emerged as a critical strategy to enhance spatial-temporal generalization and mitigate sparse training data challenges in crop yield estimations. Hybrid frameworks coupling crop growth model with machine learning algorithm, such as those proposed and evaluated by Ren et al., (2023b) and Xie and Huang, (2021), are gaining tractions.

2 Data Collection: The dataset (field measurements, meteorological/soil data, satellite imagery, crop distribution maps, and statistics) is comprehensive and representative. However, data processing steps (e.g., meteorological data interpolation, satellite image preprocessing) require more detailed technical descriptions to improve reproducibility. Furthermore, explicit clarification is needed regarding spatial alignment and scale conversion methods employed for integrating multi-resolution datasets.

Reply: Thanks for your suggestion.

We have carefully revised the Data Collection section to provide a more detailed description of the data processing procedures, particularly for the meteorological and satellite imagery data.

(1) In the revised version, we clarified the purposes and preprocessing steps for the two climate datasets (meteorological station data and climate reanalysis data) used in the study. We detailed the procedures used to address missing values and outliers in the meteorological station data. We described the resampling method employed to align the spatial resolution of ERA5 product with that of satellite imagery. (Section 2.2.2)
(2) Moreover, we expanded the description of data processing for the two satellite datasets (Sentinel-2 and MODIS LAI). We clarified that since yield maps were generated independently from each dataset for subsequent yield bias correction, we only performed reprojection to spatially align the imagery. (Section 2.2.4)

2.2.2 Meteorological data

In this study, two different climate datasets were used.

The meteorological station data used in this study came from the meteorological stations of the National Meteorological Information Center (http://data.cma.cn). There are 238 meteorological stations within the study area. Here 51 of the meteorological stations that located within 1 km buffer zone of the soybean cultivation areas were selected (Fig. 1). The meteorological datasets generally include insolation duration (h), minimum temperature (℃), maximum temperature (℃), daily average temperature (℃), average water vapor pressure (kPa), average wind speed (m sec-1), precipitation (mm) and snow-depth (cm). Observed data from 1980 to 2021 of the 51 selected stations were collected. ***Missing values and outliers in the data were filtered out. The data were then directly used for setting input climate parameters of the WOFOST model to drive simulations.***

The climate reanalysis data was obtained from the ERA5-land Daily Aggregated - ECMWF Climate Reanalysis Product. ***The data was only used to calculate soybean phenology for preparation of yield estimations.*** *It was a global climate reanalysis product that provides continuous climate data at a resolution of 0.1° × 0.1° (e.g., air temperature and atmospheric pressure) starting from 1950. The daily aggregated air temperature data at 2 m above the surface of land measured in kelvin (K) during the soybean growth periods from 2019 to 2023 was collected in this study from the Google Earth Engine (http://earthengine.google.com).* ***The product was resampled to 20 m using bilinear interpolation model to match with the resolution of satellite imagery data.***

2.2.4 Satellite imagery data

Two satellite data including: 1) Sentinel-2 Multi-Spectral Instrument (MSI) Level - 2A Surface reflectance product (10 – 60 m spatial resolution, 5-day revisit), and 2) the Moderate Resolution Imaging Spectroradiometer (MODIS) Leaf Area Index (LAI) / Fraction of Photosynthetically Active Radiation (FPAR) Level 4 product (MCD15A3H, v061, 500 m spatial resolution, 4-day period) were used to generate yield maps. All data spanning soybean growth periods (2019 – 2023) were accessed and pre-processed via the Google Earth Engine (GEE, http://earthengine.google.com).

The MSI aboard Sentinel-2A/B satellites provides 10 m (visible and near-infrared bands), 20 m (red-edge and shortwave infrared bands) and 60 m (atmospheric bands) bands at 5-day revisit. ***The Level-2A data, which are geometrically and atmospherically corrected via the Sen2Cor, were masked for clouds and shadows using the Quality Assurance (QA) band. The 60 m band was excluded due to their low spatial resolution and limited relevance for yield estimation and the 10 m (B2: Blue, B3: Green, B4: Red, B8: Near-Infrared) and 20 m (B5–B7: Red-edge, B8A: Near-Infrared, B11–B12: Shortwave Infrared) bands were retained. To harmonize spatial resolution, the 10 m bands were resampled to 20 m using bilinear interpolation model.***

The MODIS MCD15A3H (Collection 6.1, Level 4) provides 4-day composite LAI and FAPAR at 500 m derived from Terra and Aqua satellite sensors LAI/FAPAR are primarily inverted via a 3D radiative transfer model-based look-up-table (LUT) algorithm (Knyazikhin et al., 2018). When the primary algorithm fails, they are estimated using an empirical NDVI-LAI model. ***The LAI data was similarly reprojected to WGS -84 to ensure spatial alignment with Sentinel-2 imagery. These coarse-resolution LAI data were used to generate 500 m yield maps. The coarse-resolution yield maps were then used to bias-correct the 20 m Sentinel-2 yield maps, improving their regional consistency. Details about the bias correction are present in following 3.3.2 Section.***

3 Results: Results are effectively visualized through figures/tables demonstrating WOFOST model simulations, multi-scale estimation accuracy, and spatial yield patterns. The analysis appropriately discusses model accuracy, stability, and spatiotemporal pattern recognition capabilities. However, deeper interpretation of anomalies (e.g., regional/yearly estimation errors) is needed. Notably, the systematic overestimation in field-scale validation suggests potential model biases (e.g., systematic errors or overfitting), warranting further investigation.

Reply: Thanks for your suggestion.

In Result Section of the revision, we have expanded our analysis of uncertainty in soybean yield estimation at the field (Section 4.2) and regional (Section 4.3) scales. In the discussion section, we conducted a more detailed assessment of the model's estimation errors across different scales, regions, and years. The interpretation of the results is framed around two key aspects: (1) systematic errors intrinsic to WOFOST model simulations, and (2) overfitting tendences of the GRU model. On this bias, we

further discussed the limitations of the current study and suggested the directions for future research. (Section 5.3)

4.2 Yield estimation at field scale

The field-scale performance of NortheastChinaSoybeanYield20m was validated against in-situ measurement from 2022 and 2023, demonstrating strong accuracy in capturing spatial yield variability (Fig. 5). The estimated yields showed strong agreement with observed yield, with $R^2 > 0.65$ ($p < 0.01$). ***The error-bars indicated more consistent performance in fields with uniform yields, while higher uncertainties appear in fields with larger estimation deviations.*** *Overall accuracy across both years reached 0.73 in R^2 ($p < 0.01$), 287.44 kg ha^{-1} in RMSE and 10.02 % in MRE (Fig. A2). Notably, higher accuracy in 2023 with RMSE of 271.07 kg ha^{-1} and MRE of 8.57 % (Fig. 5b) was achieved. The results indicated that the dataset well captured the spatial variation of soybean yield.*

4.3.1 Variability of accuracy through years

The NortheastChinaSoybeanYield20m was validated at the municipal scale (2019 to 2022) by aggregating yield maps to match statistical data (Fig. 6). ***Compared to the field-scale validation, the municipal-scale estimates exhibited greater uncertainty, likely reflecting increased heterogeneity of soybean yields over larger areas.*** *The estimates maintained stable interannual performance, with correlation between estimated and statistical yields consistently exceeding 0.60 ($p < 0.01$). The overall accuracy, pooled across 2019- 2022, for municipal-scale achieved $R^2 = 0.62$ ($p < 0.01$), RMSE = 272.36 kg ha^{-1}, and MRE = 12.08 % (Fig. 11a). Annual accuracy metrics ranged from 221.69 kg ha^{-1} to 310.66 kg ha^{-1} for RMSE and from 8.24 % to 14.40 % for MRE, with the 2022 year achieving the highest accuracy (MRE < 10%, Fig. 6d).*

5.3 Limitations and future developments

In this study, a multi-scenario soybean growth dataset was developed by simulating various combinations input parameters within the WOFOST model. These diverse scenarios were designed to reflect different environmental and management conditions, ultimately serving as training data for the yield estimation model. One advantage of the model is its scalability, it can be readily applied to other regions and countries that lack sufficient ground observation data, such as parts of Africa and India, thus offering a promising tool for global agricultural monitoring.

However, the validation results revealed some notable limitations. Specifically, the model exhibited a tendency to produce large uncertainty in low- or high- yielding areas, introducing error into the overall yield estimation (Fig. 5 and 6). This pattern suggests a systematic bias in the model's predictions, particularly in regions with extreme yield values. Additionally, spatial analysis showed that estimation errors were more pronounced in the northern region, where is characterized by complex terrain,

compared to the relatively flat central region (Fig. 7). These discrepancies highlight the need to refine parameterization for extreme yield conditions and integrate higher-resolution environmental drivers (e.g., terrain, localized weather).

On the one hand, the estimation errors may be attributed to the inherent limitations of the WOFOST model. *As a process-based model, WOFOST simplifies its calculations for simulating physiological processes, which can hinder its ability to fully replicate the complex realities of soybean in the field. Factors, such as pest infestations, diseases, and abiotic stresses are either oversimplified or excluded (Gaso et al., 2024). These omissions can lead to systematic simulation errors, particularly under stress conditions that significantly affect crop yield. Moreover, the parameterization of the WOFOST model in this study purely relied on values from literature and existing dataset rather than local optimization. As a result, local variability because of farming practices, soil properties, and environmental conditions may not have been adequately captured. This lacks local optimization likely result in higher estimation error, especially in complex landscapes with spare ground observations. To address these issues, future works incorporating field-specific parameters or advanced data assimilation techniques could help reduce bias and improve model accuracy across heterogeneous landscapes. Given the spatial variability in soybean growth within the study area, constructing ecological zones based on factors like climate, elevation, and management practices might provide a more targeted model approach. For instance, Huang et al., (2023) defined the ecological zones through using Theissen polygons derived from meteorological station locations. This zoning strategy could enhance the representativeness of the training data and reduce yield estimation uncertainties.*

On the other hand, the estimation errors may stem from the overfitting of the GRU model. *The GRU was trained on the multi-scenarios simulated dataset, a large number of simulations that included all available combinations (e.g., all meteorological data), which introduced a significant amount of redundant information. The redundancy not only potentially reduce the dataset's representativeness, but also increase the computational burden during model training. As a result, the trained GRU model may have become overly turned to specific temporal patterns in certain years, limiting its ability to generalize to other time period or regions with different growth conditions. This overfitting effect might result in large yield estimation errors across different years and regions, particularly in areas where soybean yields deviated significantly from the norm. To address these issues, refining the structure and composition of the training dataset, and removing redundant information would enhance the diversity and quality of the training inputs. One potential approach to reduce redundancy is through spatiotemporal clustering of various environmental (e.g., meteorological station data), which could filter out stations with highly similar information. Moreover, monitoring the validation error throughout the training process, and implementing regularization techniques (e.g., L2 weight regularization) could help to prevent overfitting and improve the GRU model's generalization capability, leading to improve soybean estimation across varying conditions…*

4 Discussion: When discussing MODIS-Sentinel-2 complementarity, quantitative comparisons of their performance under varying conditions (weather/vegetation coverage) would strengthen data selection guidance. Future research directions could be expanded by aligning with emerging trends (e.g., integration with IoT/blockchain technologies, precision agriculture applications), thereby enhancing both theoretical depth and practical relevance for agricultural challenges.

Reply: Thanks for your suggestion.

(1) We compared the yield estimation performance of MODIS and Sentinel-2 under different conditions in the Discussion section (Section 5.1). Specifically, In the revised manuscript, we established 10 km grids across the study area and calculated soybean coverage of each cell. We then randomly selected three representative grid cells, corresponding to coverage thresholds of <25%, >50%, and >75%. For each selected grid cell, we extracted Sentinel-2 yield maps and MODIS LAI yield maps from 2019 to 2023 to facilitate a systematic comparison. Accordingly, the Figure 13 has been updated to quantitatively illustrate the differences between the datasets.

(2) Regarding future research directions, we have expanded our discussion on the future directions of research to explore the integration of emerging technologies such as IoT, blockchain, and precision agriculture with machine learning and biophysical models. Revisions can be found in Section 5.3.

This study generated soybean yield estimates using both MODIS LAI (500 m) products and S2 derived LAI (20 m) data. Over 2019 – 2022, the MODIS-based estimates achieved an overall R^2 of 0.58 ($p < 0.01$), an RMSE of 272.36 kg ha^{-1} and an MRE of 12.08 % (Fig. 11b), slightly lower than the Sentinel-2 based results (Fig. 11a). The uncertainty of MODIS based estimates was higher than that the Sentinel-2 based estimates, likely reflecting MODIS's coarser resolution. However, the Sentinel-2 based estimates exhibit inherent seaming effects caused by cloud-affected tile edges. We additionally used MODIS LAI to bias-correct Sentinel 2 yield maps, effectively minimizing the striping ("seaming") effects in the 20 m products (Fig. 9), while preserving pixel-level detail through tile-based calibration (Fig. 13). Despite difference in spatial resolution, both MODIS and Sentinel-2 satellite data demonstrated comparable ability to capture spatiotemporal variation in soybean yield (Fig. 12), achieving correlations with statistical data > 0.55 and overall errors < 13 % across all years.

In practical applications, balancing both temporal and spatial resolution is critical for achieving robust yield prediction results (Azzari et al., 2017). Figure 13 compares the Sentinel-2 yield maps and the MODIS LAI yield maps within a 10 km grid under different soybean coverage. Thanks to 4-day revisit, MODIS LAI provides more cloud-free observations during the critical growth stages, improving the reliability of two LAI metrics (LAI_{mean1} and LAI_{mean2}). Its coarser spatial resolution

also accelerates spatial processing over large areas. However, Sentinel-2's finer more effectively resolves intra field yield heterogeneity (Fig. 13). MODIS-derived maps occasionally underestimated yields due to mixed pixels containing non-crop features (e.g., infrastructure), whereas Sentinel-2 minimized such errors.

While this study prioritized high-resolution mapping (using MODIS solely for Sentinel-2 seam correction), combing high spatial data (e.g., Sentinel 2 or UAV imagery) with high temporal frequency satellites (e.g., geostationary sensors or radar) could provide an optimal data source for crop yield modelling (Gao and Anderson, 2019; He et al., 2018).

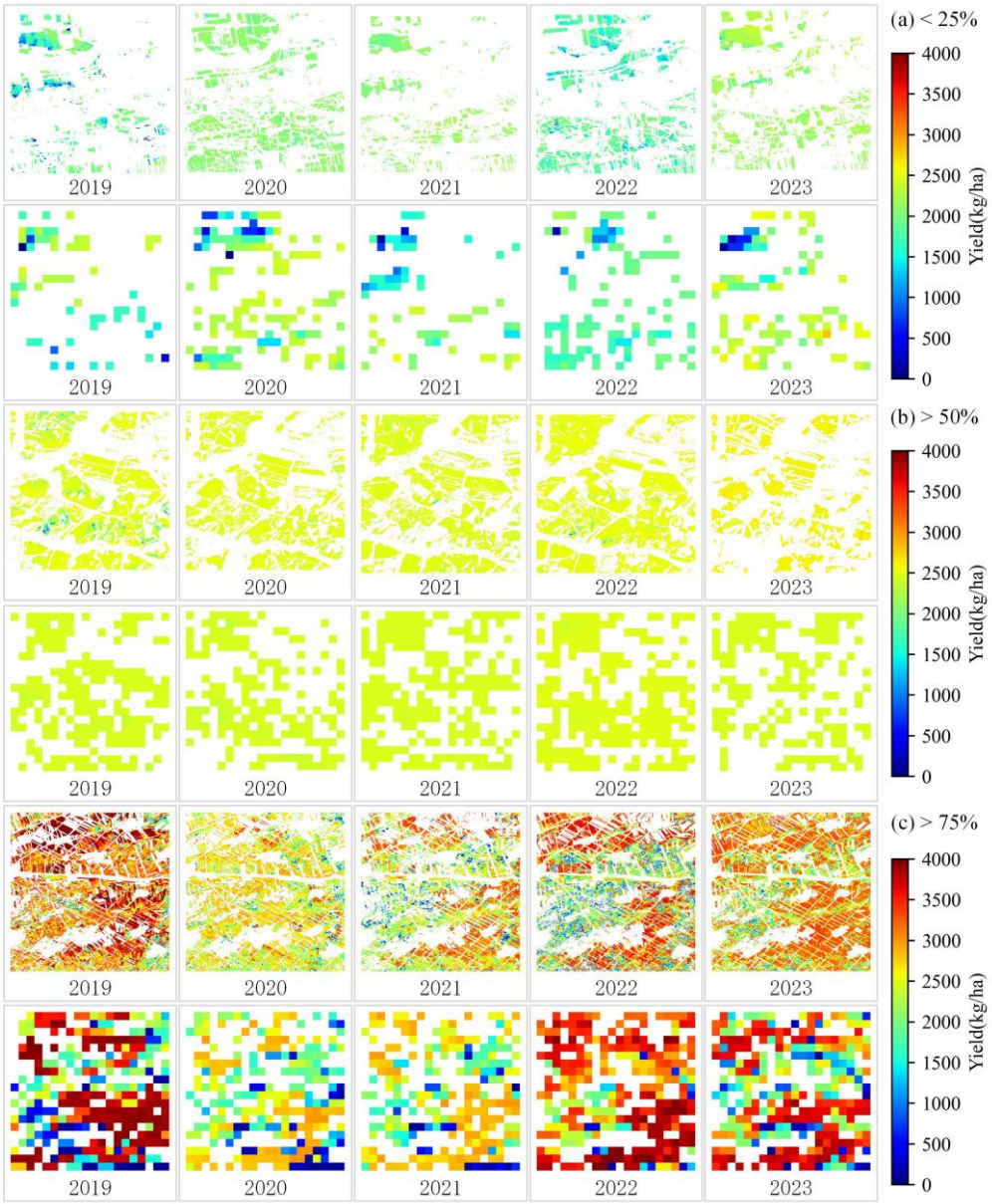

Figure 1: Comparisons of soybean yield estimation within a 10 km grid under different soybean coverage using Sentinel-2 (20 m) and MODIS LAI (500 m) data, where (a), (b), (c) represent soybean coverage less than 25%, more than 50% and more than 75%, respectively.

In addition, the combination of IoT, blockchain, and precision agriculture with machine learning and biophysical models can offer a powerful framework for sustainable agricultural monitoring, addressing challenges in data heterogeneity, model scalability, and decision-making processes. These technologies can facilitate real-time data collection, ensure data security and transparency. Precision agriculture techniques, combined with advanced sensing technologies, can effectively improve the accuracy and timeliness of input data, addressing current limitations in model calibration, validation and prediction.