Responses to the comments of Referee #3

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) The abstract was refined to emphasize the research goals, methodology, and key findings more clearly.
(2) The introduction was improved by better structuring the background information, and clearly stating the novelty of the proposed hybrid framework for soybean yield estimation.
(3) The clarity and presentation of figures were improved by enhancing the resolution and redesigning the layout.
(4) The Discussion and Conclusion section was revised to better highlight the advantages of the research and provide a more concise summary of the key findings, emphasizing the effectiveness of the proposed hybrid framework.

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.

This study presents a well-structured and logically organized framework for high-resolution soybean yield estimation. The combination of process-based modeling with deep learning offers a novel perspective for enhancing agricultural monitoring capabilities. The objectives are clearly articulated, with a strong focus on improving soybean yield data accuracy to support agricultural decision-making and production optimization. The methodological approach is rigorous, leveraging diverse production scenarios to train the GRU model and applying time-series Sentinel-2 data for large-scale yield estimation. The evaluation using in-situ measurements and government statistical data provides strong validation, and the reported accuracy metrics indicate reliable model performance across spatial and temporal scales. There are some suggestions as follows, which can be considered for further improvement of the manuscript.

Reply: Thank you for your thorough review and recognition of our work. We have carefully considered them in our revisions to enhance the quality and clarity of the manuscript.

The research is well-founded and presents significant innovations. However, the abstract and introduction sections could benefit from more professional and polished language to enhance readability and better highlight the study's contributions. Refining the writing style would improve clarity, strengthen the articulation of the research objectives, and more effectively emphasize the novelty of the proposed hybrid framework.

Reply: Thank you for your insightful comments and suggestions. In response to your valuable feedback, we have carefully refined the abstract and introduction sections to enhance the readability of the manuscript.

In the revised sections, we have improved the description of the technological background, providing a clearer discussion of data-driven and knowledge-driven approaches in crop yield estimation, along with their respective limitations. The revision ensures a more seamless transition into our research objectives and highlights the advantages of the proposed hybrid model for yield estimation. These modifications improve the overall coherence of the manuscript and better emphasize its scientific contributions.

Abstract. *Accurate monitoring of crop yield is critical for ensuring food security. While various yield datasets covering Northeast China exist, they were produced at a coarse spatial resolution and remain inadequate for capturing small-scale spatial heterogeneity. Current yield estimation methods, such as machine learning models and the assimilation of remotely sensed biophysical variables into crop growth models, are heavily reliant on ground observations and computationally expensive. To address these limitations, we propose a hybrid framework that couples the World Food Studies Simulation Model (WOFOST) and a Gated Recurrent Unit (GRU) model to generate*

a high-resolution (20 m) soybean yield dataset in Northeast China from 2019 to 2023 (NortheastChindaSoybeanYield20m). First, to generate a comprehensive training dataset, WOFOST was employed to simulate diverse soybean growth scenarios by accounting for variations in climates, crop varieties, soil types and agro-managements practices. The GRU model was then trained to establish relationships between model simulated leaf area index (LAI) and soybean yield. The trained model was applied to estimate soybean yield in Northeast China using time-series LAI derived from Sentinel-2 at key growth stages. The accuracy of estimates was evaluated using in-situ measurements and government statistical data. The overall accuracy was 287.44 kg ha^{-1} and 272.36 kg ha^{-1} in the root mean squared error (RMSE) for field and regional scale, respectively. The model exhibited consistent interannual stability, with mean relative error (MRE) averaging 11.46 % and 7.94% at the municipal scale and the provincial scale, respectively. The dataset effectively captured spatiotemporal yield variability, offering potentials for optimizing soybean production, guiding precise agriculture practices, and informing agricultural policy. The NortheastChinaSoybeanYield20m dataset is publicly available at https://doi.org/10.5281/zenodo.14263103 (Xu et al., 2024).

1 Introduction

Soybean is a crucial crop for both food and oil production, providing more than a quarter of the world's edible protein (Graham and Vance, 2003). Global demand for soybean is projected to increase by 46 % by 2050, driven by rapid population growth (Falcon et al., 2022). As an major traded agricultural commodity, soybean production in key exporting nations has wide-reaching effects on international markets, and can significantly influence agricultural economies worldwide (Qiao et al., 2023). Notably, China is the world's largest consumer of soybeans (FAOSTAT, 2022), and its soybean demand relies heavily on international trade (Zhao et al., 2023). Consequently, accurate monitoring of soybean yield is vital for promoting sustainable agriculture, ensuring food security, and maintaining economic stability from regional to global scale. Moreover, effective yield monitoring and mapping supports farmers by informing field management practices, bolstering agricultural insurance and enhancing poverty alleviation initiatives (Zhuo et al., 2022).

Remote sensing data provides time-series observations for crop yield estimation across multiple scales (e.g., field, regional and national) (Dong et al., 2020; Hunt et al., 2019; Zhao et al., 2023b). Current methodologies for yield estimation can be broadly categorized as data-driven or knowledge-driven approaches.

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. These approaches utilized simulated outputs from crop growth models (e.g., meteorological, soil, crop physiological, and management factors) as inputs for machine learning, reducing reliance on limited ground observations Many studies have demonstrated hybrid methods are able to enhance yield estimation due to three benefits (Feng et al., 2020; Xie and Huang, 2021; Yang et al., 2021). The simulations from crop growth model can provide biophysical constraints to machine learning, ensuring agronomic plausibility. The crop growth models generate synthetic training datasets to address data scarcity. Finally, the machine learning improves the computational efficiency compared to traditional data assimilation techniques (Xie and Huang, 2021). However, exiting studies generally extracted input features (e.g.,

LAI, and soil moisture) across the entire growth cycle or on coarse temporal scales, increasing computational costs of model calculation and obscuring stage-specific physiological response (Pinke and Lövei, 2017; Wang et al., 2015). Additionally, while deep learning models, such as Long Short-Term Memory (LSTM) and GRU model excel at modelling temporal dependencies, their integration into hybrid frameworks have not been widely explored.

Critically, the primary soybean-producing regions of China lack a publicly available high-resolution yield dataset to analyse spatiotemporal production patterns, hindering precision agriculture and policy optimization. To address this, we developed a hybrid model coupling the World Food Studies (WOFOST) crop growth model with a GRU deep learning method to estimate soybean yield in Northeast China. The objectives include: (1) Design a hybrid framework integrating WOFOST-simulated growth scenarios with GRU-based temporal feature extraction; (2) Generate a high-resolution (20 m) soybean yield dataset in Northeast China (NortheastChinaSoybeanYield20m) from 2019 to 2023; (3) Evaluate the accuracy of the dataset across field, municipal, and provincial scales using in situ and statistical benchmarks. The WOFOST model first simulated a multi-scenario soybean growth (varying climate, soil, crop varieties and management conditions) to train the GRU model. The time series Sentinel-2 data, capturing soybean growth development, were then input into the GRU model to estimate yield. This approach prioritizes stage-specific physiological dynamics which balancing computational efficiency and spatial granularity, providing a critical advancement for scalable agricultural monitoring.

Figure 1: where is the soybean classification map from? What is the accuracy?

Reply: Thank you for the comments.

The soybean map in Figure 1 was derived from existing study of Zhao et al., (2022) using an optimal identification feature (OIF) knowledge graph coupled with a moment-preserving segmentation method. The study classified maize, soybean, and rice in the Northeast China. The soybean distribution maps from 2019 to 2023 were collected in this study. The overall accuracy and the producer accuracy for maize, soybean and rice was higher than 90 % and 93 %, respectively, with a Kappa coefficient greater than 0.90.

In the revision, details on soybean classification maps were presented in Section 2.2.5. We have added additional information to Figure. 1, including the source of the classification map and classification accuracy.

Figure 1: Location of the study area and the distribution of sample plots in two years (2022 and 2023) and selected meteorological stations. The soybean distribution map was obtained from Zhao et al., (2022) using a moment-preserving segmentation method, achieving an overall accuracy over 90% for soybean in 2023 (Details are provided in Section 2.2.5).

2.2.5 Crop distribution data

The soybean distribution maps for the study area (2019 – 2023) were obtained from Zhao et al., (2022), which employed a novel methodology for crop type identification. The study proposed an optimal identification feature (OIF) knowledge graph coupled with a moment-preserving segmentation method to classify crop types without ground-truth data. The method achieved overall accuracy above 90% and producer's accuracy exceeding 93% for maize, soybean and rice, with a Kappa coefficient greater than 0.90.

Zhao, L., Li, Q., Chang, Q., Shang, J., Du, X., Liu, J., and Dong, T.: In-season crop type identification using optimal feature knowledge graph, ISPRS Journal of Photogrammetry and Remote Sensing, 194, 250–266, https://doi.org/10.1016/j.isprsjprs.2022.10.017, 2022.

Figure 5 appears blurry, which affects the clarity and readability of the presented data. I suggest organizing box plots and histograms as subfigures.

Reply: Thanks for your suggestion. In response, we have reorganized the histograms (a) and the box plots (b) as subfigures to present the data more effectively. We believe these adjustments improve the visualization and overall presentation of the results.

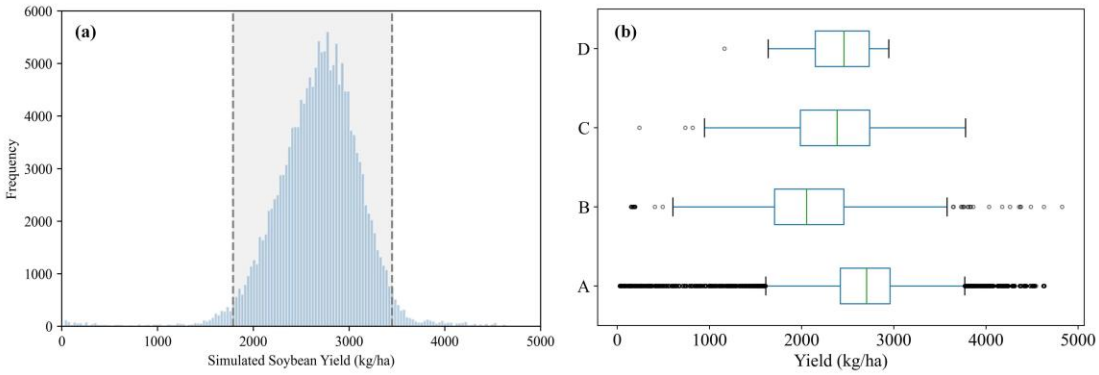

Figure 2: (a) Histogram statistics of simulated soybean yield where the gray area in the histogram represents 95 % confidence intervals; (b) distribution of simulated soybean yield compared with other datasets where A represents simulated yield in this study (n = 171,360), B represents statistical yield from 1980 to 2022 (n = 961), C represents specific measurements from the literature (Chen et al., 2011; Fan et al., 2012; Liu et al., 2005, 2008; Liu and Herbert, 2002; Wang et al., 2020, 2024; Zheng and Zhang, 2021) (n = 138) and D represents measurements in 2022 and 2023 carried by this study (n = 39).

The discussion on the advancements of the proposed method is embedded within the "Limitations and future developments" section. To better highlight the strengths of this study, I recommend extracting this content into a standalone subsection. This

would allow for a clearer and more structured presentation of the method's advantages, making it easier for readers to appreciate its contributions in comparison to existing approaches.

Reply: Thank you for your suggestion.

(1) In the revised version, we have introduced Section 5.2 "Advancements in this Study" to provide a clearer discussion of the method's advantages. We have improved the logical coherence and academic professionalism of the content in Discussion to enhance readability.
(2) Furthermore, we have included a comparative analysis between our research findings and existing methodologies, which better demonstrates the superiority of our approach in terms of accuracy, computational efficiency, and large-scale applicability.

Below is the revised content:

5.2 Advancements in this study

Accurate monitoring of soybean yield is crucial for food policy decision-making and security assessment. While previous studies have primarily explored the impact of environmental factors such as climate on soybean productivity (Guo et al., 2022; Zhao et al., 2023a), few efforts have focused on producing high-resolution soybean yield dataset for China's major soybean-producing regions. To address this gap, our study produced the NortheastChinaSoybeanYield20m dataset, a 20-meter resolution dataset generated through a hybrid framework integrating the mechanistic WOFOST crop growth model and a GRU deep learning algorithm. Unlike purely data-driven approaches that rely on extensive ground data, our approach leveraged both data mining capabilities and mechanistic modelling, which improve the model's interpretability and enhances its potential for transferability across regions. The integration of the WOFOST model ensured the simulation of diverse production scenarios under varying climate, soil, crop variety and management conditions, providing a robust synthetic training data for the GRU network. This combination allowed the model to generate well, even in areas with limited observational data, therefore overcoming common limitations related to data scarcity and high computational costs. Accuracy assessments using both in-situ and statistical yield data confirmed that the generated NortheastChinaSoybeanYield20m dataset delivered reliable yield estimates across field and regional scales (Fig. 5 and 6). The results also verified the model's stability across time and space, reinforcing its potential for large-scale agricultural monitoring and strategic planning.

When compared to previous studies using integrated remote sensing data and process-based model to estimate soybean yield, for instance, Baup et al., (2015) reported estimation error ranging from 2% to 18%, our method achieved comparable levels of accuracy. It also outperformed existing field-scale studies (e.g., RMSE = 400.946 kg ha^{-1} in Ren et al., (2023) and MRE of 29.73% in Du et al., (2014)) and

municipal-scale models (e.g., RMSE = 16 % in Von Bloh et al., (2023)). Furthermore, the NortheastChinaSoybeanYield20m dataset showed improved performance relative to similar high-resolution soybean yield products from other countries (e.g., annual 30 m soybean yield mapping in Brazil, with R^2 values between 0.31 and 0.71 and RMSEs ranging from 275 to 740 kg ha^{-1} (Song et al., 2022).

Although studies based on UAV and RGB data have demonstrated even higher soybean yield estimation accuracy (Li et al., 2021, 2024), such methods are often constrained by high costs and limited spatial coverage, making them impractical for large-scale applications. In contrast, the method developed in this study offers a well-balanced solution that combines computational efficiency, high spatial resolution, and strong predictive accuracy. Our approach offers scalable and practical solution for producing high-resolution, large-scale crop yield datasets.

The conclusion effectively summarizes the study but could be further refined to better highlight the innovation in dataset construction and its practical applications in agricultural management.

Reply: Thank you for your valuable feedback. In the revised version, we have enhanced the conclusion to emphasize the novel aspects of our approach, particularly the integration of the WOFOST model with deep learning, as well as the practical implications of the NortheastChinaSoybeanYield20m dataset for agricultural management.

Here is the revised conclusion.

This study generated a high-resolution (20 m) soybean yield dataset for Northeast Chinda from 2019 to 2023 (NortheastChinaSoybeanYield20m) using a hybrid framework that couple the WOFOST crop growth model with a Gated Recurrent Unit (GRU) deep learning algorithm. The framework leveraged a comprehensive soybean growth dataset simulated by WOFOST, which accounted for diverse production scenarios, including variations in climates, crop varieties, soil types and agro-managements practices. This approach effectively reduces reliance on ground observation data, which demonstrating enhanced spatiotemporal generalization capabilities.

The dataset was conducted using multi-source remote sensing data, with Sentinel-2 derived time-series LAI as the primary input. Yield estimations showed robust performance at both field and municipal scales, achieving RMSE of 287.44 kg ha^{-1} and 272.36 kg ha^{-1}, respectively. To address spatial discontinuities in Sentinel-2 data, corrections using MODIS LAI-derived yield maps effectively mitigated seam effects, achieving complementary benefits in temporal and spatial resolution. The final dataset exhibits high temporal stability and spatial continuity, with mean relative errors (MRE) averaging of 11.46 % at the municipal scale and 7.94 % at the provincial scale.

The NortheastChinaSoybeanYield20m dataset successfully captures fine-scale

spatiotemporal variations in soybean yield, offering potentials for optimizing production strategies, guiding precision agriculture, and enhancing food security and policy.