Dear Editor,

I am writing to submit a revised version of our manuscript (essd-2026-284), entitled "**NortheastChinaMaizeYield10m: A 10-m Resolution Maize Yield Dataset for Northeast China (2019–2024) Generated via a Mechanistically Interpretable, Label-free Framework**", originally submitted to "Earth System Science Data" on 14 Apr 2026. We appreciate the opportunity to revise our work in response to the reviewers' and editors' feedback.

We have carefully considered and addressed each comment provided by the reviewers in the revision. In particular, we have refined the terminology "label-free" to "field-label-free training" to avoid conceptual ambiguity, added LSTM as a baseline model to justify our architectural choice, incorporated SHAP analysis to strengthen mechanistic interpretability, and expanded independent field validation with newly collected 2025 measurements. We have also added systematic comparisons with existing public yield products, included county-level validation at a finer administrative scale, provided uncertainty layers and bootstrap confidence intervals for model performance metrics, and expanded the discussion on the simulation-to-reality gap.

We have attached a detailed response letter outlining the changes made in response to each comment from the reviewers. All edits in the attached revised manuscript are marked in red color for your assessments. A clean version of the revision is also attached. We hope that this demonstrates our commitment to improving the manuscript in line with the journal's standards and requirements.

We confirm that all authors have approved this revised version of the manuscript and agreed to its submission to "Earth System Science Data" for reconsideration. We believe that our study makes a valuable contribution to the field and is well-suited for publication in your esteemed journal.

We deeply appreciate your consideration of our revised manuscript for review. We sincerely welcome any critiques, comments, and suggestions for improving the manuscript. Should you have any questions, please don't hesitate to let us know.

Yours sincerely,
Xin Du

Corresponding author:
Dr. Xin Du
Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
Tel./Fax: +86-010-6483-7645
E-mail: duxin@aircas.ac.cn

Responses to the comments of Referee #1

Article ID: essd-2026-284

Title: NortheastChinaMaizeYield10m: A 10-m Resolution Maize Yield Dataset for Northeast China (2019–2024) Generated via a Mechanistically Interpretable, Label-free Framework

Authors: Jingbo Hu, Xin Du, Qiangzi Li, Yuan Zhang, Hongyan Wang, Jiansong Luo, Jingyuan Xu, Yachao Zhao, Zhaoming Zhang, Yong Dong, and Yunqi Shen

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). **Terminological refinement.** We have replaced the term "label-free" with "field-label-free" throughout the manuscript, and added a dedicated statement in the Introduction to precisely define its meaning and distinguish it from conventional supervised and unsupervised learning paradigms.

(2). **Enhanced data preprocessing descriptions.** We have expanded the technical details for Sentinel-2 imagery preprocessing in the Data Sources section, and clearly documented the sources and acquisition of the official statistical datasets.

(3). **Clarification of LAI as the core feature.** We have added a dedicated discussion to elucidate the physiological rationale for selecting LAI as the central input feature, articulating its role as an integrated proxy for canopy structure, photosynthetic capacity, and environmental stress responses, as well as its advantages for large-scale crop monitoring.

(4). **Strengthened discussion on simulation–reality discrepancies.** We have supplemented Section 4.4 to explain the model's capacity to simulate waterlogging-induced stress, and significantly extended Section 4.5 to enumerate the key uncertainties and limitations arising from input data deviations and systematic differences between simulated and satellite-derived LAI.

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 manuscript presents a well-structured and comprehensive framework for large-scale maize yield estimation by integrating process-based modeling, remote sensing data, and a GRU deep learning model. The overall framework is well designed and addresses the common challenge of limited field yield observations in large-scale agricultural monitoring. The generation of a 10 m resolution maize yield dataset covering Northeast China from 2019 to 2024 further demonstrates the practical value of the framework. The evaluation based on both in-situ measurements and government statistical data provides strong validation. The study is relevant to the scope of the journal and provides useful insights into the integration of process-based models and deep learning for yield estimation.

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 The term "label-free" is used frequently in the article, from the title to the main text. However, the study still relies on WOFOST-simulated labels during model training. It would be helpful for the authors to clarify the exact meaning and scope of "label-free" in the introduction, particularly how it differs from conventional supervised and unsupervised learning methods.

Reply： Thank you very much for your insightful and constructive comment.

We completely agree with the reviewer that using the absolute term "label-free" without precise contextual boundaries could lead to conceptual confusion, as the Gated Recurrent Unit (GRU) model does rely on the yield outcomes simulated by the WOFOST model as target labels during its training phase.

To eliminate any potential ambiguity and ensure strict terminological rigor, we have thoroughly revised the manuscript. Specifically, we have consistently replaced the term "label-free" with "field-label-free training" throughout the title, abstract, introduction, and Methods section. This revision explicitly clarifies that our framework is completely free of real-world / in situ ground yield annotations during the model training loop.

In addition, following your valuable suggestion, we have added a dedicated statement in the Introduction to clearly define the exact meaning and scope of “field-label-free” and to clarify its essential differences from conventional machine learning paradigms.

To address these challenges, this study proposes a field-label-free training framework for maize yield estimation that couples mechanistic model with deep learning.

To address the aforementioned challenges, this study constructs a field-label-free training framework for large-scale maize yield estimation.

From a practical application perspective, this study not only constructs a low-cost, scalable, mechanism-driven, field-label-free training estimation solution…

Unlike conventional supervised learning, which relies on real-world ground-truth labels, or unsupervised learning, which does not involve target variables, the proposed framework employs process-based crop model simulations to generate synthetic target labels.

2 The dataset (field measurements, meteorological/soil data, satellite imagery, crop distribution maps, and statistics) is comprehensive and representative. However, some data preprocessing steps require more detailed technical descriptions. For example, the pre-processing for Sentinel-2 imagery is only briefly introduced, and the source and acquisition of the statistical datasets are not clearly documented.

Reply： Thank you for your recognition of our dataset and your constructive feedback.

We have significantly expanded the technical descriptions in the Data Sources section (Section 2.2), specifically addressing the two points you raised.

2.2.1 Remote Sensing Data

Sentinel-2 imagery acquired by the Multi-Spectral Instrument (MSI) was selected as the primary remote sensing data source for retrieving leaf area index (LAI) in this study. Sentinel-2 provides relatively high spatial and temporal resolution data (10 m and 20 m spatial resolution with a five day revisit interval) and includes three red-edge spectral bands, which offer distinct advantages for monitoring vegetation conditions during the middle and late stages of crop growth (Drusch et al., 2012) . Using the Google Earth Engine (GEE) cloud computing platform, Sentinel-2 Level-2A surface reflectance products covering the study area were collected for the maize growing seasons from 2019 to 2024, spanning from April to October. These products have been radiometrically calibrated and atmospherically corrected. Notably, because the time series reconstruction algorithm employed in this study can utilize all valid pixels, no cloud filtering was performed at the initial collection stage. To eliminate interference from clouds and shadows, we applied cloud masking to all images using the QA60 band. The 60-m bands were excluded due to their low spatial resolution and limited relevance for field-scale yield estimation. In contrast, the 10-m bands (B2: Blue, B3: Green, B4: Red, B8: Near-Infrared) and key 20-m red-edge bands (B5 and B8A) were retained.

2.2.6 Statistical data

Official annual maize yield statistics at the county level from 2019 to 2024 were collected from the Statistical Yearbooks publicly released by the Statistical Bureaus of Heilongjiang Province (http://tjj.hlj.gov.cn), Jilin Province (http://tjj.jl.gov.cn), Liaoning Province (https://tjj.ln.gov.cn), and the Inner Mongolia Autonomous Region (https://tj.nmg.gov.cn). However, because the Statistical Yearbooks data were not published for certain regions in some years, the available data cover only a subset of regions in those specific years. These statistical data are primarily used to validate the results of this study, and are also combined with the published yield records to assess *the reasonableness of the simulated yields produced by our approach.*

3 The current input features are almost exclusively centered around the LAI. Some important factors affecting yield formation, such as water stress and extreme temperature conditions, are not explicitly considered. It is suggested that the authors further elucidate the rationale behind ultimately selecting LAI as the core feature for modeling. Especially, the authors may discuss whether LAI can be regarded as an integrated proxy of crop canopy growth and photosynthesis accumulation, as well as the advantages of using LAI for large-scale yield estimation.

Reply： Thank you for this exceptionally profound and constructive comment.

We fully agree that water stress and extreme temperature events are critical factors affecting crop yield formation, and we acknowledge that they are not explicitly included as direct input features in our current modeling framework. We have added a dedicated discussion in the Discussion (4.3 Physiological Interpretation of LAI Features) to explain why LAI was selected as the core feature.

As the core input feature, the LAI is not merely a static structural variable; rather, it serves as a comprehensive characterization of both canopy architecture and photosynthetic capacity, which directly determines the fraction of intercepted photosynthetically active radiation and consequently drives biomass accumulation (Parker, 2020). The temporal dynamics of LAI throughout the entire growing season can reflect the cumulative impacts of water stress and extreme temperature conditions, as any environmental disturbances will inevitably leave distinct imprints on the seasonal LAI sequence through inhibited leaf expansion, accelerated senescence, or changes in canopy greenness (Chen et al., 2025; Zhang et al., 2025). Furthermore, LAI can be stably retrieved from high-resolution satellite imagery such as Sentinel-2, making it suitable for large-scale, high-frequency crop monitoring, whereas observational datasets for moisture and temperature are often difficult to acquire at comparable spatial and temporal scales. Therefore, selecting LAI as the core feature in this study successfully balances mechanistic interpretability with data feasibility.

4 The discussion regarding the differences between simulated and real-world observations could be further strengthened. While Figures 4 and 5 demonstrate the overall representativeness of the simulated data, some factors in real agricultural systems are not explicitly represented in the WOFOST simulations. Additional discussion of these potential limitations would improve the rigor of the manuscript.

Reply： Thanks for your suggestion.

We acknowledge that the fundamental premise of our framework—training a deep learning model entirely on simulated data and transferring it to real-world satellite observations—relies on the representativeness of the WOFOST simulations. While our validation results confirm the feasibility of this transfer, several inherent gaps between the simulated and real domains warrant explicit discussion. Accordingly, we have supplemented the manuscript with Section 4.4, which explains the model's capacity to simulate waterlogging-induced stress, and have significantly extended Section 4.5 to enumerate the key discrepancies between simulated and real domains.

4.4 Simulation and Model Response to Extreme Precipitation Stress

In this study, the WOFOST model was operated under the water-limited production mode, which is not equivalent to simulating only drought stress. The WOFOST soil water balance module is driven by actual daily precipitation and dynamically tracks changes in daily soil moisture content (van Diepen et al., 1989). Although the WOFOST model does not explicitly simulate the full range of physiological processes associated with anaerobic root conditions under waterlogging, this simplification does not prevent the model from capturing waterlogging-induced yield losses. When cumulative precipitation exceeds the soil drainage and runoff capacity, the model inherently triggers excess-water stress responses—suppressing root oxygen availability, limiting canopy transpiration and photosynthesis, and ultimately reducing LAI and biomass accumulation(H. et al., 1998; Liu et al., 2020). Therefore, our training dataset inherently includes yield reduction scenarios caused by excessive soil moisture, which originate directly from the 30-year historical meteorological records (1995–2024). Sentinel-2 satellite observations are capable of capturing such LAI anomaly signals, and the GRU model has learned, during the training phase, the general mapping relationship between depressed LAI trajectories and reduced final yields under various environmental stresses. When applied to real satellite imagery, the model does not explicitly detect flood events; instead, it infers yield losses through the observed LAI decline as a comprehensive stress signal.

4.5 Uncertainties and Limitations

Although this study achieved relatively high yield estimation accuracy, the model still carries certain uncertainties due to data source limitations and simplified assumptions. The gap between crop growth model simulations and reality, along with input data uncertainties, are the primary sources constraining the reliability of crop yield estimates.

The input layer of the WOFOST simulations, including both cultivar parameters and agronomic management settings, inevitably deviates from the complexity of real world conditions. In this study, several key physiological parameters governing phenological development were assigned based on literature values and the model's default settings, rather than calibrated against site-specific field trials for each cultivar. In terms of agronomic management, our framework assumes non-limiting nutrient conditions and discretizes the sowing window into only four fixed dates. In reality, smallholder farmers across Northeast China employ flexible sowing schedules and highly diverse management strategies (Chuanwei et al., 2024). These two types of input deviations *may jointly cause the simulated LAI trajectories to diverge to some extent from actual growth curves in specific fields. Nevertheless, we argue that the full-factorial parameter combination strategy adopted in this study constructs an exhaustive training set that covers a broad physiological and management space. The consistency between estimated yields and statistical yearbook data (Figure 11), together with the robust multi-year independent field validation performance (Section 3.4), demonstrates that the simulated data generated under the current input settings are sufficiently reliable to support accurate yield estimation.*

A systematic discrepancy also exists between the WOFOST-simulated LAI and the satellite- derived LAI. The simulated values represent a deterministic, noise-free "green leaf area index" at the point scale under idealized atmospheric conditions (H. et al., 1998). In contrast, the satellite-retrieved LAI is subject to multiple factors, including residual atmospheric correction errors, cloud-gap filling algorithms, and mixed-pixel effects. As noted by Lu et al. (2025), satellite-derived vegetation parameters (e.g., LAI) are not free from uncertainty, particularly in areas with high LAI or dense canopies, where saturation or underestimation may occur. Moreover, different data processing workflows, such as cloud masking and filtering algorithms, can introduce systematic errors. Consequently, input uncertainties can propagate through the prediction model when a model trained on clean simulated data is generalized to noisy satellite observations (Xie et al., 2025). However, the gating mechanism of the GRU provides inherent robustness to such input noise by dynamically weighting the contribution of each time step (Cho, 2014). The stable performance of the model in independent validation across 2022–2024 (Table 7) further confirms that the proposed framework can effectively mitigate the adverse effects of uncertainties in satellite-derived LAI under the current data quality.

5 The discussion regarding the ability of the simulated dataset to represent extreme conditions may be somewhat overstated. Although the results from the 2023 disaster year demonstrate the potential robustness of the framework under adverse conditions, they may not be sufficient to prove comprehensive coverage of all extreme scenarios. A more cautious interpretation of these findings is recommended.

Reply： Thank you for your valuable comments.

We agree that a more cautious interpretation is needed regarding the model's performance under extreme conditions. Accordingly, we have added a dedicated discussion in Section 4.5 ("Simulation and Model Response to Extreme Precipitation Stress") to more carefully address this issue and provide a balanced assessment of the model's capabilities and limitations.

4.4 Simulation and Model Response to Extreme Precipitation Stress

In this study, the WOFOST model was operated under the water-limited production mode, which is not equivalent to simulating only drought stress. The WOFOST soil *water balance module is driven by actual daily precipitation and dynamically tracks changes in daily soil moisture content* (van Diepen et al., 1989). *Although the WOFOST model does not explicitly simulate the full range of physiological processes associated with anaerobic root conditions under waterlogging, this simplification does not prevent the model from capturing waterlogging-induced yield losses. When cumulative precipitation exceeds the soil drainage and runoff capacity, the model inherently triggers excess-water stress responses—suppressing root oxygen availability, limiting canopy transpiration and photosynthesis, and ultimately reducing LAI and biomass accumulation(H. et al., 1998; Liu et al., 2020). Therefore, our training dataset inherently includes yield reduction scenarios caused by excessive soil moisture, which originate directly from the 30-year historical meteorological records (1995–2024). Sentinel-2 satellite observations are capable of capturing such LAI anomaly signals, and the GRU model has learned, during the training phase, the general mapping relationship between depressed LAI trajectories and reduced final yields under various environmental stresses. When applied to real satellite imagery, the model does not explicitly detect flood events; instead, it infers yield losses through the observed LAI decline as a comprehensive stress signal.*

Responses to the comments of Referee #2

Article ID: essd-2026-284

Title: NortheastChinaMaizeYield10m: A 10-m Resolution Maize Yield Dataset for Northeast China (2019–2024) Generated via a Mechanistically Interpretable, Label-free Framework

Authors: Jingbo Hu, Xin Du, Qiangzi Li, Yuan Zhang, Hongyan Wang, Jiansong Luo, Jingyuan Xu, Yachao Zhao, Zhaoming Zhang, Yong Dong, and Yunqi Shen

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). **Added 2025 independent field validation.** We collected new in situ measurements from the 2025 growing season, which lies entirely outside the WOFOST simulation period. The model achieved an R^2 of 0.61 and RMSE of 1.65 t/ha on this unseen year, confirming genuine temporal generalization. (Section 3.4, Figure 9)

(2). **Strengthened baseline comparisons by incorporating LSTM models.** To better justify our architectural choice, we implemented LSTM networks under the optimal feature configuration (Scheme C). The LSTM model showed substantial performance degradation when transferred to real-world conditions (R^2 = 0.31), confirming GRU's superior generalization from simulated to real-world data. (Section 3.2.2, Table 7, Appendix Figure A2)

(3). **Clarified the rationale for full-factorial simulation design.** We clarified that our dataset has undergone quality control based on model simulation feasibility and physical plausibility, and the strong independent validation performance (R^2 = 0.69, RMSE = 1.21 t/ha) confirms that the training data support robust transferable learning. (Section 2.4.2)

(4). **Strengthened validation of 2023 disaster impact using county-level spatial analysis.** We conducted county-level yield anomaly analysis for 2023 relative to the multi-year average (2019–2024) and compared the spatial distribution of estimated yields against reported flood-affected areas (China Climate Bulletin, 2023; (Duo et al., 2024; Yin et al., 2025)). The affected counties exhibit substantially lower yields compared to adjacent unaffected counties , providing independent spatial evidence that our product effectively captures flood-induced yield anomalies. (Section 3.5, Figure 12)

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 offers substantial data value and promising application potential, and the availability of ground-based validation data is particularly valuable. However, the manuscript requires further improvement. My specific comments are as follows.

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.

Specific Comments:

1 The authors randomly shuffled the WOFOST-generated samples and divided them into training, validation, and test subsets at an 8:1:1 ratio. However, these samples were generated from a finite set of meteorological stations, years, soil types, crop parameter combinations, and sowing dates. Under a purely random split, samples sharing the highly similar parameter settings are likely to appear in both the training and test sets. Consequently, the reported test performance may primarily reflect interpolation within a previously observed simulation space rather than genuine generalization to unseen years, locations, or climatic conditions. I therefore recommend that the authors adopt grouped or out-of-distribution evaluation strategies, such as leave-one-year-out, leave-one-station-out, spatial block holdout, or joint station–year holdout experiments.

Reply： We sincerely thank the reviewer for this rigorous and insightful comment regarding the potential data leakage issue in random sample splitting.

To address this concern, we have now expanded our independent field validation to include the newly collected 2025 growing season, which lies entirely outside the WOFOST simulation period (1995–2024) used for training data generation. This provides a rigorous out-of-sample temporal test of the model's generalization capability to completely unseen climatic conditions. The 2025 validation alone yielded an R^2 of 0.61 and RMSE of 1.65 t/ha, confirming that the model genuinely generalizes to unseen years rather than merely interpolating within the simulation space.

The updated validation results, including the 2025 field measurements, have been incorporated into the revised manuscript in **Section 3.4 and Figure 9**.

To further evaluate the model's generalization to completely unseen climatic conditions, we extended the field validation to include the newly collected 2025 growing season using the optimal GRU model (Scheme C). The 2025 season lies entirely outside the WOFOST simulation period (1995–2024) used for training data generation, providing a rigorous out-of-sample temporal test. As shown in Figure 9, the model achieved an R^2 of 0.61 and RMSE of 1.65 t/ha on the 2025 validation set, consistent with the performance on 2022–2024 (R^2 = 0.69, RMSE = 1.21 t/ha). This confirms that the model genuinely generalizes to unseen years rather than merely interpolating within the simulation space.

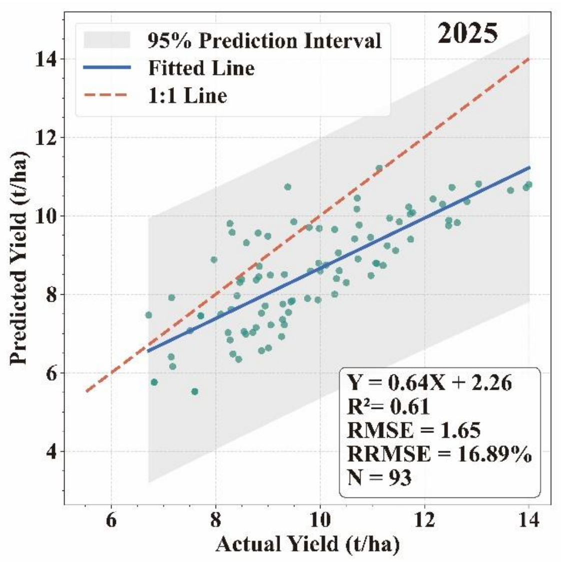


Figure 9 Independent validation of the optimal GRU model (Scheme C) against in situ measurements in 2025.

2 The study assumes that meteorological conditions, soil types, crop parameters, and sowing dates can be combined independently. In reality, these factors are strongly interdependent. Although a full-factorial design broadens the simulated parameter space, it may also produce trajectories that are numerically valid in WOFOST but agronomically implausible. The authors should therefore impose region-specific suitability constraints on parameter combinations, weight simulated samples using observed cultivar zoning, sowing-date surveys, or agricultural statistics. Clearly unrealistic combinations should be excluded before model training.

Reply： We sincerely thank the reviewer for this thoughtful comment regarding the independence of parameter combinations in our full-factorial design. We acknowledge that in reality, meteorological conditions, soil properties, crop varieties, and sowing dates are strongly interdependent. However, we would respectfully argue that our current design is justified for the following reasons.

First, our dataset has already undergone a quality control procedure based on model simulation feasibility and physical plausibility. We excluded simulation runs that failed to complete due to model convergence issues or produced non-physiological results, such as failure to reach maturity within the growing season, zero or negative biomass accumulation, or biologically implausible LAI trajectories. The 3,686,400 valid samples reported in our study represent the dataset after applying these necessary filters, ensuring that only physiologically plausible growth trajectories are retained for model training.

Second, and most importantly, our study is fundamentally a transfer learning task from simulation to reality, and the true test of our training data lies in the independent field validation. The GRU model achieves strong generalization performance across 2022–2024 (R^2 = 0.69, RMSE = 1.21 t/ha) and maintains consistent accuracy on the 2025 validation set. If the training dataset contained a substantial number of agronomically implausible trajectories that negatively biased the model, we would expect poor performance when transferred to real-world satellite observations. The consistent accuracy across multiple years—including the extreme flood year of 2023 and the entirely unseen year of 2025—provides compelling empirical evidence that our training data, despite being generated through full-factorial combinations, have enabled the model to learn robust and transferable mapping relationships between growth dynamics and final yield.

Rather than being a limitation, the diversity of parameter combinations is a key strength that allows the model to generalize across the wide range of real-world conditions captured in the independent validations. We have clarified this rationale in the revised manuscript.

By systematically combining long-term meteorological records from 60 stations spanning 30 years, four major soil texture types, four representative sowing dates, and extensive crop parameter variations (192 crop parameter combinations), a maize growth simulation dataset with pronounced spatiotemporal heterogeneity was constructed using the PCSE framework. Our full-factorial design prioritizes physiological state space coverage over exhaustive enumeration of all possible agronomic combinations, ensuring that the training dataset encompasses a wide range of growth rates, phenological durations, and stress responses under Northeast China's climatic and edaphic conditions. In total, 3,686,400 valid simulation samples were generated, and the detailed parameter combinations are summarized in Table 3. This dataset preserves complete time series trajectories of leaf area index (LAI) under diverse environmental stress conditions, together with the corresponding final grain yield, thereby providing a comprehensive crop growth simulation dataset for subsequent deep learning surrogate model training.

3 Although RF is a reasonable static baseline, outperforming RF does not by itself demonstrate that GRU effectively captures the continuous and cumulative nature of crop growth. The authors should include a broader set of baselines, including state-of-art models.

Reply: Thank you for this excellent and practical suggestion. To better justify our architectural choice and demonstrate the suitability of GRU for this task, we have expanded our baseline comparisons by incorporating the LSTM model under our primary feature configuration (Scheme C).

The quantitative performance comparison between GRU and LSTM has been incorporated into **Section 3.2 (Comparative Performance of Deep Learning Models)**

3.2.2 LSTM Models

As shown in Table 7, the LSTM model achieves moderate performance on the simulation test set (R^2 = 0.57, RMSE = 0.96 t/ha) but exhibits substantial degradation when transferred to real-world conditions, with an independent validation R^2 of only 0.31 (RMSE = 1.67 t/ha). This substantial gap suggests that GRU's more compact gating mechanism, with fewer parameters, offers superior generalization from simulated training data to noisy real-world satellite observations. Given its competitive performance and higher computational efficienc, GRU was selected as the primary model for this study. The complete LSTM results, including scatter plots of model performance on training, test, and validation sets, along with SHAP feature importance analysis, are provided in Appendix Figure A2

Table 7 Performance of LSTM models under Scheme C (Full Features) and independent validation using in situ yield measurements (2022–2024)

Scheme ID	**Accuracy**	**LSTM model performance on simulation dataset**			**Independent validation using measured yield**			
		Training set	Test set	Validation set	2022	2023	2024	All years combined
C (Full Features)	R^2	0.61	0.57	0.27	0.31	0.4	0.24	0.31
	RMSE (t/ha)	0.91	0.96	1.25	1.77	1.52	1.84	1.67
	RRMSE (%)	10.06	10.6	13.85	20.65	16.82	21.22	18.99

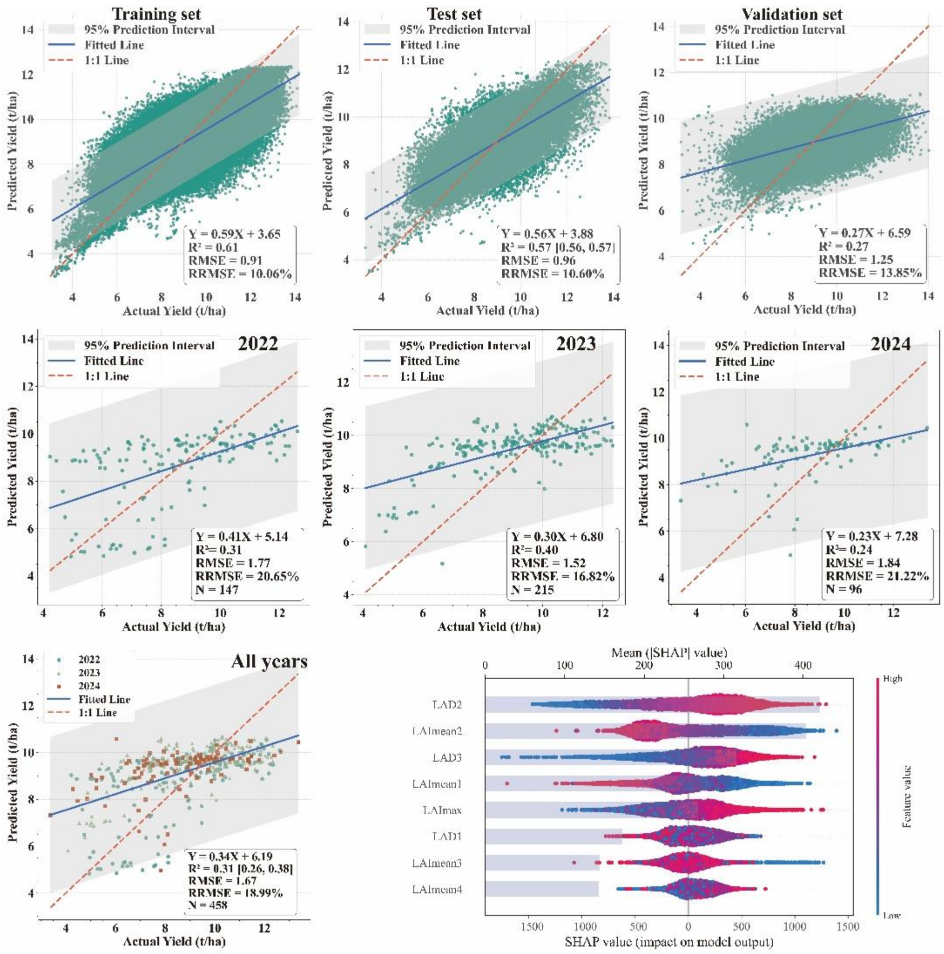


Figure A2. Performance of LSTM models under Scheme C (Full Features) and independent validation using in situ yield measurements (2022–2024)

4 The proposed product successfully captured yield anomalies caused by the 2023 typhoon and flooding events. I recommend that the authors incorporate independent disaster-related reference data, such as officially reported affected areas or remote-sensing-derived flood extent products, and compare yield anomalies between disaster-affected areas and adjacent unaffected areas.

Reply: We sincerely thank the reviewer for this constructive suggestion regarding the validation of disaster-related yield anomalies. To address this comment, we collected officially reported information on the flood-affected areas during the 2023 extreme precipitation events in Northeast China (China Climate Bulletin, 2023; (Duo et al., 2024; Yin et al., 2025)) and compared the yield anomalies between affected and adjacent unaffected regions. Specifically, we generated two complementary figures: (a) the county-level yield anomaly map showing the relative deviation of 2023 yields from the multi-year average (2019–2024), calculated as ($yield_{2023} - yield_{mean}$) / $yield_{mean}$; and (b) the spatial distribution of 2023 county-level estimated yields across the study region. As shown in the newly added **Figure 12**, the affected areas exhibit substantially lower yields compared to adjacent unaffected regions, with the most pronounced reductions concentrated in areas consistent with the officially reported flood zones. This spatial correspondence provides independent evidence that our product effectively captures the yield anomalies induced by the 2023 flooding events.

Notably, the generated yield maps successfully captured a pronounced regional yield reduction in 2023 (Figure 10). In contrast to the relatively homogeneous high-yield patterns observed in normal years such as 2022 and 2024, the yield estimates for 2023 exhibited widespread low-value zones across most parts of the study area, particularly in the central and southern regions. To quantitatively assess the spatial pattern of this anomaly, we computed county-level yield anomalies for 2023 relative to the multi-year average (2019–2024), defined as $(yield_{2023} - yield_{mean}) / yield_{mean}$*. As shown in Figure 12(a), the most severe negative anomalies are concentrated in southern Heilongjiang Province and northern Jilin Province. In contrast, adjacent areas outside the reported flood zones exhibit near-normal or only moderately reduced yields, confirming the localized nature of the disaster impact. We further compared the spatial distribution of 2023 county-level estimated yields against the officially reported flood-affected zones (China Climate Bulletin, 2023; (Duo et al., 2024; Yin et al., 2025)). As shown in Figure 12(b), the counties within the documented flood-affected areas—including Wuchang, Shangzhi, Yushu, and Shulan—exhibit substantially lower yields (generally 6.0–8.0 t/ha) compared to adjacent unaffected counties (typically 8.5–10.0 t/ha). This spatial correspondence between low-yield zones and the flood extents documented in official reports and previous studies provides independent evidence that our product effectively captures the yield anomalies induced by the 2023 flooding events.*

Figure 11 (A, B, C) illustrate the spatial distribution of maize yield in the severely affected area along the boundary between Wuchang City and Yushu City before, during, and after the disaster year. In non disaster years such as 2022 and 2024, maize yields in this region were generally maintained at high levels ranging from 9.5 to 11.0 t ha^{-1}*, reflecting its strong production potential as a core area within the "golden maize belt." In contrast, during the disaster year of 2023, the estimated yields declined sharply to approximately 7.0–9.0 t* ha^{-1}*. Early August coincides with the critical phenological stage from tasseling and silking to early grain filling, during which maize is particularly sensitive to water conditions. Prolonged waterlogging stress caused root hypoxia and decay, which subsequently accelerated leaf chlorosis and premature senescence in the aboveground canopy. This process substantially reduced photosynthetic assimilation and ultimately led to significant yield losses.*

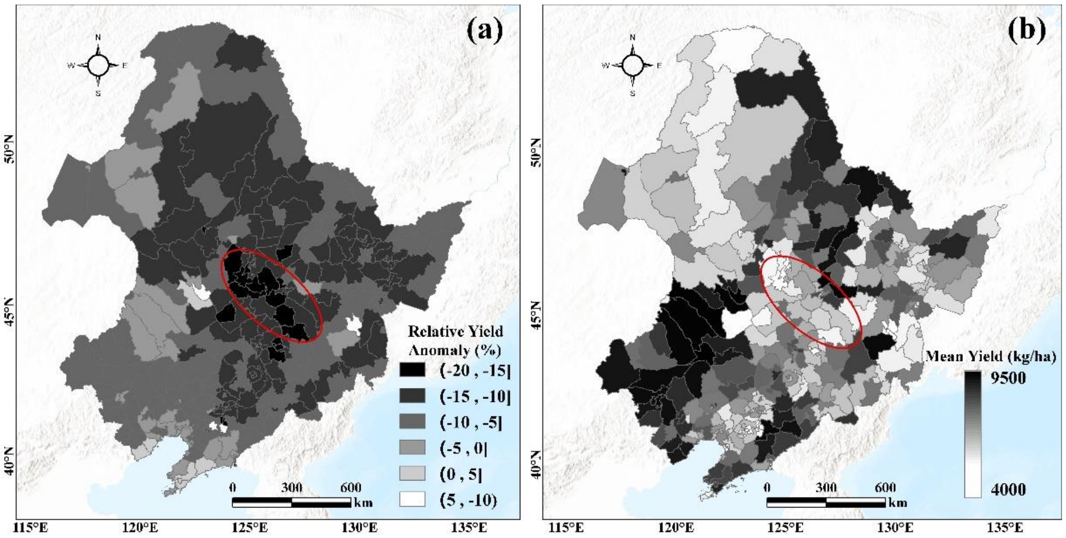


Figure 12 County-level analysis of the 2023 flood-induced maize yield anomalies in Northeast China. (a) Spatial distribution of county-level yield anomalies in 2023 relative to the multi-year average (2019–2024). (b) Spatial distribution of estimated county-level maize yields in 2023. The most severe anomalies concentrated in southern Heilongjiang and northern Jilin provinces.

Responses to the comments of zhao zhang

Article ID: essd-2026-284

Title: NortheastChinaMaizeYield10m: A 10-m Resolution Maize Yield Dataset for Northeast China (2019–2024) Generated via a Mechanistically Interpretable, Label-free Framework

Authors: Jingbo Hu, Xin Du, Qiangzi Li, Yuan Zhang, Hongyan Wang, Jiansong Luo, Jingyuan Xu, Yachao Zhao, Zhaoming Zhang, Yong Dong, and Yunqi Shen

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). **Refined the terminology "label-free" to "field-label-free training".** We have consistently replaced this term throughout the manuscript to precisely clarify that our framework is free of real-world in situ yield annotations during model training.

(2). **Strengthened baseline comparisons by incorporating LSTM models.** We implemented LSTM networks under the optimal feature configuration to better justify our architectural choice.

(3). **Enhanced mechanistic interpretability through SHAP analysis.** We conducted SHAP analysis on the optimal GRU model to quantify the contribution of each feature to yield prediction.

(4). **Provided uncertainty layers and bootstrap confidence intervals.** We generated 10-meter resolution uncertainty maps (CV) for 2023 and 2024, which are now released alongside the yield product. Due to the intensive computational workload associated with pixel-wise uncertainty quantification across our large-scale study area (approximately 1.24×10^6 km^2) and the current data download and processing constraints of the Google Earth Engine (GEE) platform, we have currently completed and released the uncertainty layers for 2023 and 2024 as a first priority. The remaining years (2019–2022) are under active production and will be released progressively upon completion. We also conducted bootstrap resampling to report 95% confidence intervals for all deep learning model configurations in Appendix A.

(5). **Added comparison with existing public yield products.** We compared our product with the published "A 10 m maize, rice and soybean yield dataset from 2016 to 2021 in Northeast China" (Teng et al., 2026) through both quantitative and visual assessments.

(6). **Added county-level validation.** We collected official county-level maize yield statistics for 2019–2024 and conducted additional validation at this finer administrative scale.

(7). **Expanded discussion on simulation-to-reality gap and clarified simulation design rationale.** We added dedicated discussion on uncertainties and clarified that our full-factorial design prioritizes physiological state space coverage with quality-controlled samples.

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.

I find the general idea of this study quite innovative. In particular, the use of crop-model simulations to generate training samples, combined with remote-sensing time-series data for high-resolution yield estimation, has clear methodological value. However, I believe several aspects of the manuscript need to be further strengthened to improve its rigor, interpretability, and the practical usability of the resulting data product.

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.Clarify the use of the term "Label-free".

The manuscript currently presents "Label-free" as a central contribution, but this term may be somewhat misleading. I suggest that the authors define it more precisely, for example as "field-label-free training," "simulation-driven training," or "simulation-driven yield estimation." This distinction should be clearly stated in the abstract, introduction, and methods sections to avoid the impression that the entire study is completely free of any field observations.

Reply： Thank you very much for your insightful and constructive comment. We completely agree with the reviewer that using the absolute term "label-free" without precise contextual boundaries could lead to conceptual confusion, as the Gated Recurrent Unit (GRU) model does rely on the yield outcomes simulated by the WOFOST model as target labels during its training phase.

To eliminate any potential ambiguity and ensure strict terminological rigor, we have thoroughly revised the manuscript. Specifically, we have consistently replaced the term "label-free" with "field-label-free training" throughout the title, abstract, introduction, and Methods section. This revision explicitly clarifies that our framework is completely free of real-world / in situ ground yield annotations during the model training loop.

Specifically, we have added a dedicated statement in the Introduction to clearly define the exact meaning and scope of “field-label-free”.

To address these challenges, this study proposes a field-label-free training framework for maize yield estimation that couples mechanistic model with deep learning.

To address the aforementioned challenges, this study constructs a field-label-free training framework for large-scale maize yield estimation.

From a practical application perspective, this study not only constructs a low-cost, scalable, mechanism-driven, field-label-free training estimation solution…

2.Add comparisons with existing yield products.

The manuscript would benefit from a more systematic comparison with existing public yield products, such as SPAM and the “A 10 m maize, rice and soybean yield dataset from 2016 to 2021 in Northeast China” dataset. Both visual comparisons and quantitative accuracy assessments would help demonstrate the added value of the proposed product. In addition, the reported accuracy appears to be lower than that of current existing machine-learning-based yield estimation studies. The authors should discuss the possible reasons for this difference, such as the simulation-to-reality gap, limited field observations, simplified management assumptions, or uncertainties in remote-sensing inputs.

Reply： Thank you for this constructive comment and for highlighting these two important aspects. We appreciate the opportunity to clarify our comparative evaluation and the underlying sources of yield estimation uncertainty. Below, we address both points in detail:

1. Comparison with Existing Public Yield Products

We fully agree that comparing our product with established regional yield datasets enhances the credibility and demonstrates the added value of our map.(Teng et al., 2026)

In the revised manuscript, we have acquired the published *"A 10 m maize, rice and soybean yield dataset from 2016 to 2021 in Northeast China"* dataset. We conducted a systematic quantitative validation and cross-comparison between our product, this public dataset, and official county-level statistical yearbooks for overlapping years.

3.5.2 Comparison with Existing Public Yield Products

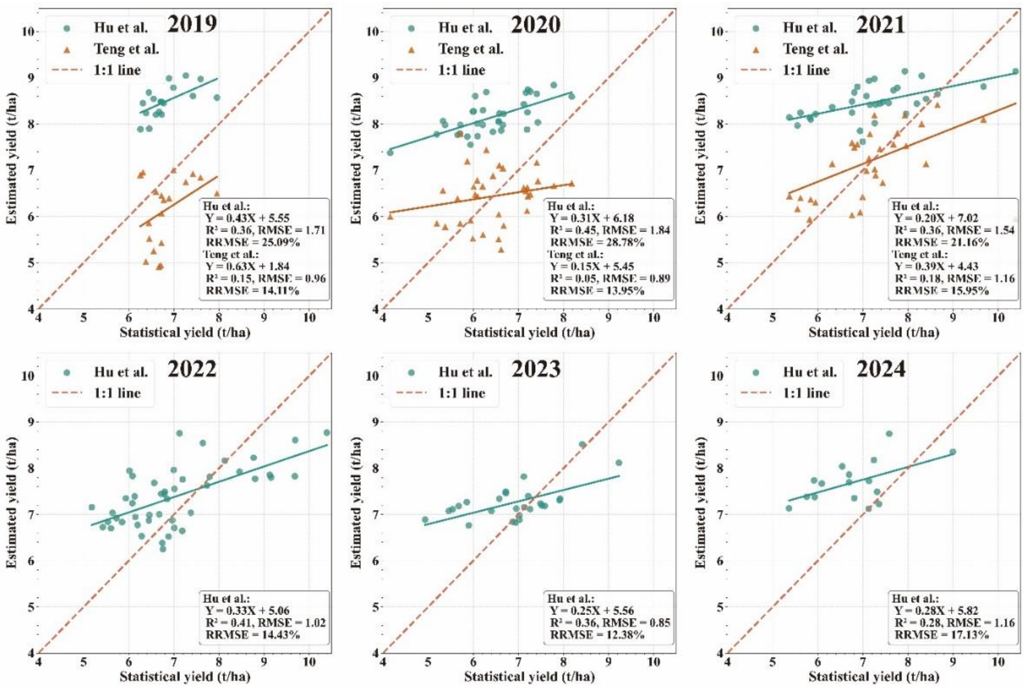


Figure 14 County-level comparison of maize yield estimates between our product and the existing 10-m benchmark dataset** (Teng et al., 2026) **against official statistical yearbook data in Northeast China (2019–2021).

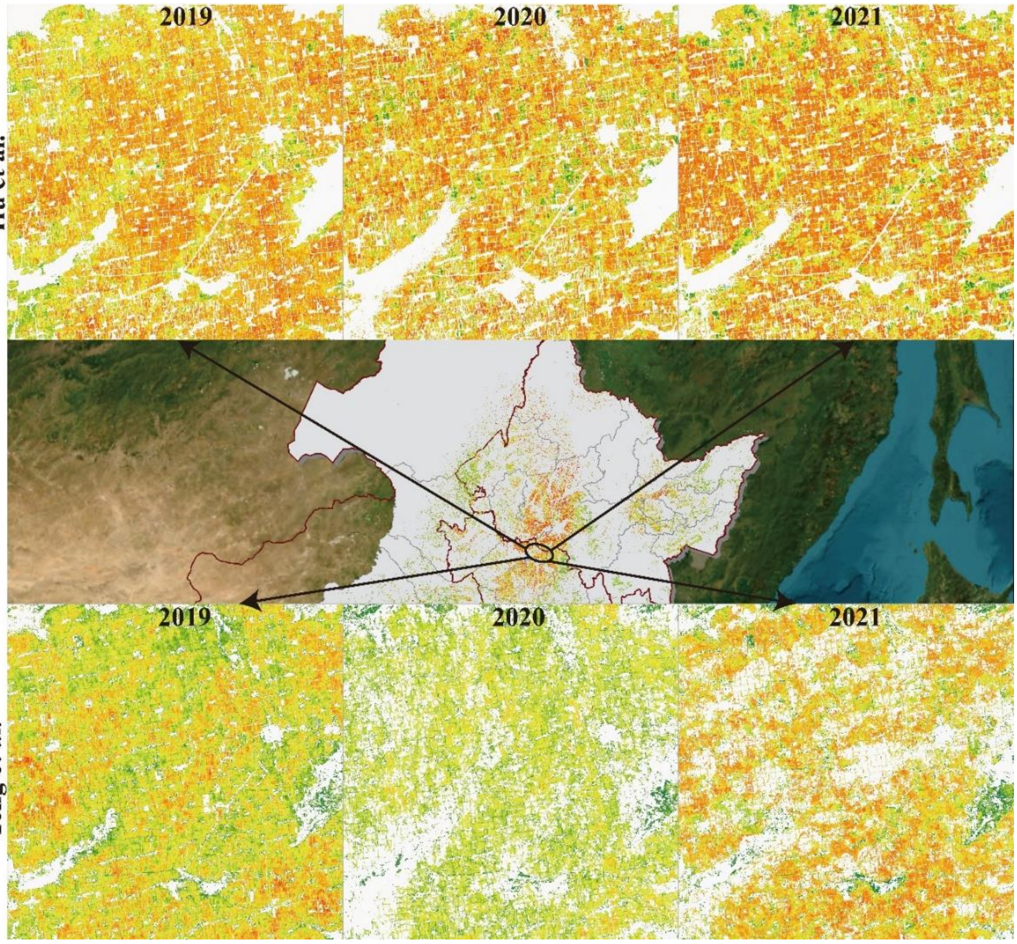


Figure 15 Visual comparison of spatial maize yield mapping between our product and the existing 10-m benchmark dataset (Teng et al., 2026) ***in the major maize-producing areas at the junction of Heilongjiang and Jilin provinces.***

We compared our product with the publicly available " A 10 m maize, rice and soybean yield dataset from 2016 to 2021 in Northeast China "(Teng et al., 2026) *over the overlapping period (2019–2021). Both products were evaluated against official county-level statistical yearbook data (Figure 12).*

As shown in Figure 14, our product achieves consistently higher agreement with county-level statistics across all three years compared to the benchmark dataset. At the county-level (Figure 14), our product demonstrates robust performance across the full 2019–2024 period, with R^2 values ranging from 0.28 to 0.45 and RRMSE from 12.38% to 17.13%.

Beyond quantitative accuracy, we also assessed the spatial depiction quality of the two products through visual comparison. Figure 15 presents the spatial distribution of estimated maize yields in the major maize-producing area at the junction of Heilongjiang and Jilin provinces. Our product exhibits clearly delineated field boundaries and more spatially continuous within-field patterns, with fewer fragmented or missing pixels. Non-crop linear features such as field ridges are effectively removed, resulting in cleaner and more homogeneous field parcels. Our product achieves favorable accuracy against the existing benchmark and provides added value through extended temporal coverage (2019–2024), improved spatial accuracy at finer administrative scales, and enhanced field-level spatial detail.

2.Possible reasons affecting the accuracy of the report

Regarding the point on reported accuracy, we have dedicated **Section 4.5 (Uncertainties and Limitations)** to thoroughly discuss the primary factors contributing to estimation errors.

4.5 Uncertainties and Limitations

Although this study achieved relatively high yield estimation accuracy, the model still carries certain uncertainties due to data source limitations and simplified assumptions. The gap between crop growth model simulations and reality, along with input data uncertainties, are the primary sources constraining the reliability of crop yield estimates.

The input layer of the WOFOST simulations, including both cultivar parameters and agronomic management settings, inevitably deviates from the complexity of real world conditions. In this study, several key physiological parameters governing phenological development were assigned based on literature values and the model's default settings, rather than calibrated against site-specific field trials for each cultivar. In terms of agronomic management, our framework assumes non-limiting nutrient conditions and discretizes the sowing window into only four fixed dates. In reality, smallholder farmers *across Northeast China employ flexible sowing schedules and highly diverse management strategies* (Chuanwei et al., 2024). *These two types of input deviations may jointly cause the simulated LAI trajectories to diverge to some extent from actual growth curves in specific fields. Nevertheless, we argue that the full-factorial parameter combination strategy adopted in this study constructs an exhaustive training set that covers a broad physiological and management space. The consistency between estimated yields and statistical yearbook data (Figure 11), together with the robust multi-year independent field validation performance (Section 3.4), demonstrates that the simulated data generated under the current input settings are sufficiently reliable to support accurate yield estimation.*

A systematic discrepancy also exists between the WOFOST-simulated LAI and the satellite- derived LAI. The simulated values represent a deterministic, noise-free "green leaf area index" at the point scale under idealized atmospheric conditions (H. et al., 1998)*. In contrast, the satellite-retrieved LAI is subject to multiple factors, including residual atmospheric correction errors, cloud-gap filling algorithms, and mixed-pixel effects. As noted by Lu et al.* (2025)*, satellite-derived vegetation parameters (e.g., LAI) are not free from uncertainty, particularly in areas with high LAI or dense canopies, where saturation or underestimation may occur. Moreover, different data processing workflows, such as cloud masking and filtering algorithms, can introduce systematic errors. Consequently, input uncertainties can propagate through the prediction model when a model trained on clean simulated data is generalized to noisy satellite observations* (Xie et al., 2025)*. However, the gating mechanism of the GRU provides inherent robustness to such input noise by dynamically weighting the contribution of each time step* (Cho, 2014)*. The stable performance of the model in independent validation across 2022–2024 (Table 7) further confirms that the proposed framework can effectively mitigate the adverse effects of uncertainties in satellite-derived LAI under the current data quality.*

3.Include additional model-structure comparisons.

To better justify this model choice, I suggest adding other time-series models, such as LSTM and Transformer-based models, as baseline comparisons. This would help show whether GRU is indeed the most suitable architecture for this task, rather than simply being one possible option.

Reply： Thank you for this constructive suggestion. To better justify our architectural

choice and demonstrate the suitability of GRU for this task, we have expanded our baseline comparisons by incorporating the LSTM model under our primary feature configuration (Scheme C).

The quantitative performance comparison between GRU and LSTM has been incorporated into **Section 3.2 (Comparative Performance of Deep Learning Models)**

3.2.2 LSTM Models

*As shown in Table 7, the LSTM model achieves moderate performance on the simulation test set (R^2 = 0.57, RMSE = 0.96 t/ha) but exhibits substantial degradation when transferred to real-world conditions, with an independent validation R^2 of only 0.31 (RMSE = 1.67 t/ha). This substantial gap suggests that GRU's more compact gating mechanism, with fewer parameters, offers superior generalization from simulated training data to noisy real-world satellite observations. Given its competitive performance and higher computational efficienc, GRU was selected as the primary model for this study. The complete LSTM results, including scatter plots of model performance on training, test, and validation sets, along with SHAP feature importance analysis, are provided in **Appendix Figure A2***

Table 7 Performance of LSTM models under Scheme C (Full Features) and independent validation using in situ yield measurements (2022–2024)

Scheme ID	**Accuracy**	**LSTM model performance on simulation dataset**			**Independent validation using measured yield**			
		Training set	Test set	Validation set	2022	2023	2024	All years combined
C (Full Features)	R^2	0.61	0.57	0.27	0.31	0.4	0.24	0.31
	RMSE (t/ha)	0.91	0.96	1.25	1.77	1.52	1.84	1.67
	RRMSE (%)	10.06	10.6	13.85	20.65	16.82	21.22	18.99

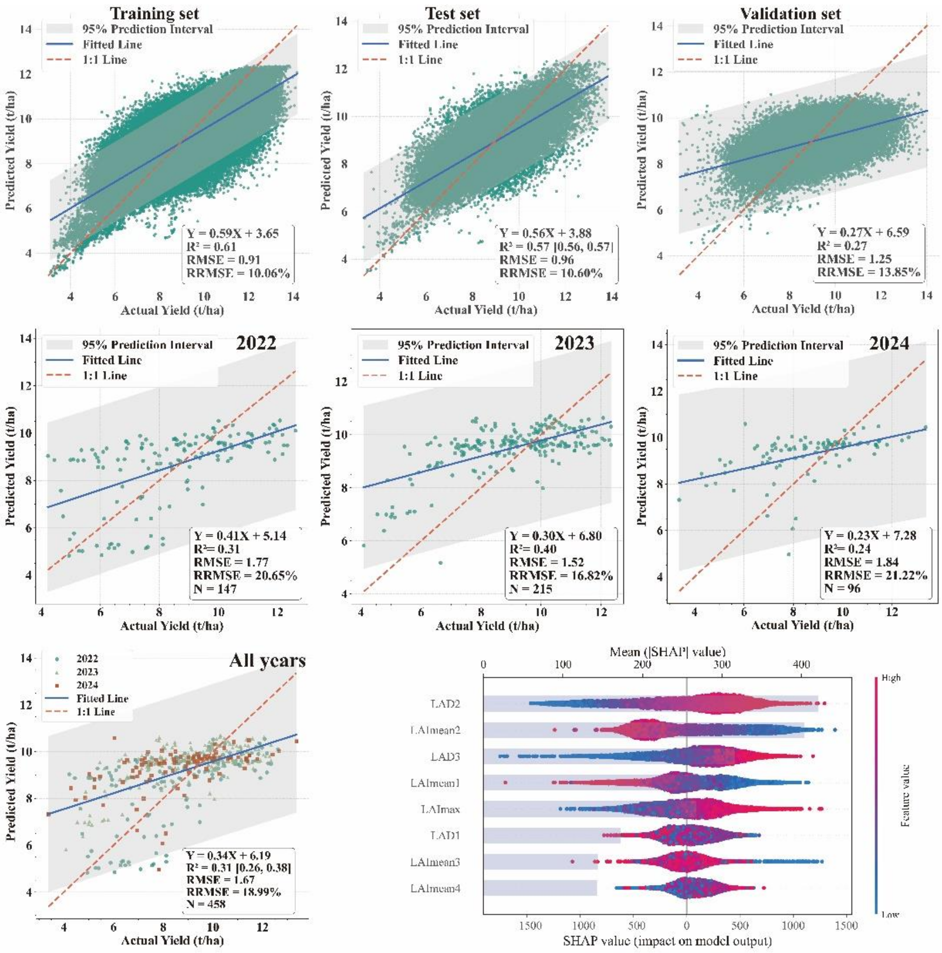


Figure A2. Performance of LSTM models under Scheme C (Full Features) and independent validation using in situ yield measurements (2022–2024)

4.Strengthen the mechanistic interpretability analysis.

The manuscript mainly relies on feature-combination experiments to assess the effects of different input variables. I think it is not sufficient to support a strong claim of "mechanistic interpretability" .I suggest incorporating model interpretation tools such as SHAP to quantify the dynamic contributions of different factors across growth stages to the final yield estimation. This would provide a clearer understanding of how the model uses information from different phenological periods and environmental variables.

Reply: Thank you for this highly valuable and insightful suggestion. In accordance with your recommendation, we have incorporated SHAP (SHapley Additive exPlanations) analysis into our revised manuscript to evaluate feature contributions dynamically. Specifically, we applied SHAP to both our primary GRU model and the newly added LSTM baseline. The complete SHAP interpretability analysis has been incorporated into **Section 3.2.1 (GRU Models)** of the revised manuscript.

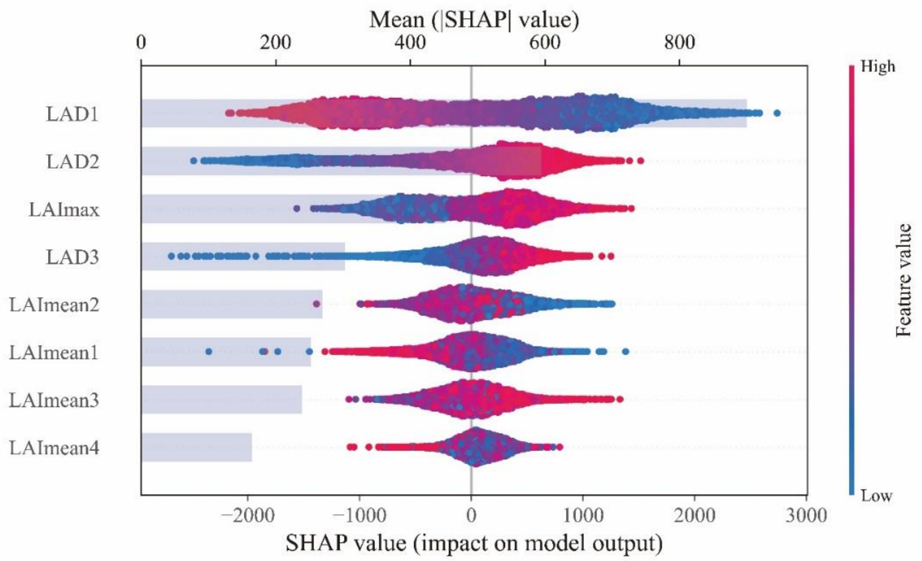


Figure 7 SHAP feature importance analysis for the optimal GRU model (Scheme C)

To provide a quantitative assessment of individual feature contributions, we conducted a SHAP analysis on the optimal GRU model (Scheme C). SHAP provides a unified measure of feature importance by quantifying each input's marginal contribution to the model output across all possible feature subsets. Figure 7 presents the SHAP summary plot for all eight features in Scheme C, with features sorted by their mean absolute SHAP values in descending order. The results reveal that cumulative features (LAD_1–LAD_3) dominate the model's decision-making, collectively accounting for 56.61% of *the total SHAP importance. LAD_1 (Emergence to Flowering) exhibits the highest individual contribution at 28.35%, followed by LAD_2 (Flowering to Milking Maturity) at 18.72% and LAD_3 (Milking Maturity to Maturity) at 9.54%. The peak feature LAI_{max} ranks third with 14.25% contribution, indicating that maximum canopy development is an important indicator of yield potential, though secondary to cumulative growth metrics. This implies that achieving a high peak LAI is not sufficient; the temporal accumulation of photosynthetic capacity across the season is more decisive.*

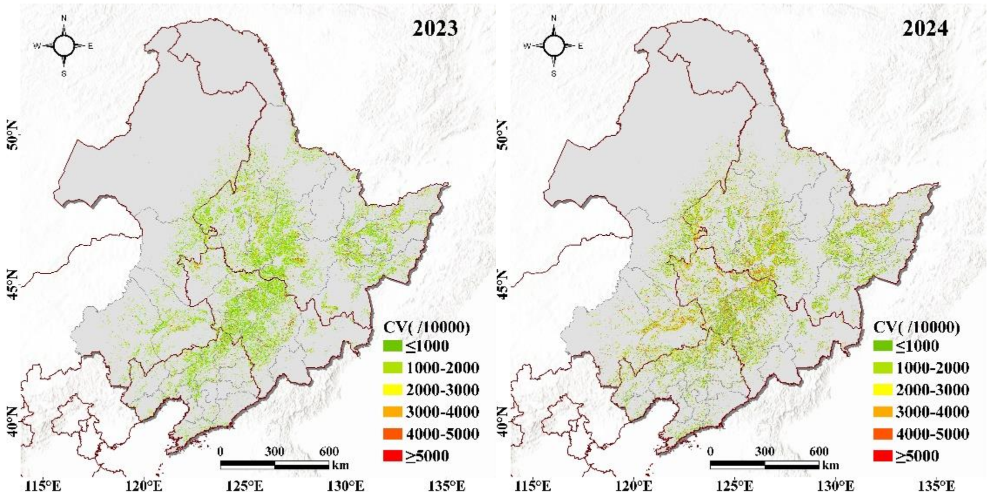
5.Provide confidence or uncertainty layers for the yield product.

For a data product intended for practical use, a single yield estimate may not be enough. I suggest providing annual 10 m confidence or uncertainty maps in TIF format. Such layers would allow users to identify where the yield estimates are more reliable and where they should be interpreted with caution.

Reply： Thank you for this excellent and practical suggestion. We completely agree that providing spatial confidence or uncertainty layers is crucial for data users to evaluate the reliability of local yield predictions and apply the dataset effectively.

In response to this comment, we have generated annual 10-meter resolution yield uncertainty maps based on Monte Carlo Dropout (MC-Dropout), a widely adopted approach for uncertainty quantification in deep learning models. By performing multiple stochastic forward passes with dropout layers activated during inference, we obtain a distribution of predictions for each pixel, from which we derive the standard deviation and the coefficient of variation (CV = standard deviation / mean). The CV represents the relative prediction uncertainty, providing a normalized metric that facilitates cross-regional comparisons. Due to the intensive computational workload associated with pixel-wise uncertainty quantification across our large-scale study area (approximately 1.24×10^6 km²) and the current data download and processing constraints of the Google Earth Engine (GEE) platform, we have currently completed and released the uncertainty layers for 2023 and 2024 as a first priority. These two years cover both a normal year (2024) and an extreme flood year (2023), providing users with uncertainty information under contrasting climatic conditions. The remaining historical years (2019–2022) are under active production and will be publicly released as soon as processing completes. We have added a clear description of the uncertainty layer availability in the Data availability section of the revised manuscript. The CV values in the uncertainty layers have been scaled up by a factor of 10,000 for integer storage; users should divide the pixel values by 10,000 to obtain the original CV values.

The detailed description has been incorporated into the revised manuscript in the Data availability section.



10-meter resolution uncertainty maps (CV) for 2023 and 2024(True value = pixel value / 10000)

The Northeast China Maize Yield 10-m dataset (2019–2024) is openly available at https://zenodo.org/records/19547014 (Hu et al., 2026). Accompanying 10-meter resolution uncertainty layers (coefficient of variation, CV) for 2023 and 2024 are also provided; uncertainty layers for 2019–2022 are under active production and will be released progressively upon completion. The CV layers were generated using Monte Carlo Dropout, with CV values scaled by 10,000 for integer storage; users should divide pixel values by 10,000 to recover the original CV. Users are encouraged to consult these uncertainty layers to assess the spatial reliability of the yield estimates.

6.Report uncertainty in model-performance metrics.

Machine-learning models are affected by randomness in model training, sample splitting, and parameter initialization. Therefore, I suggest reporting uncertainty ranges or confidence intervals for the main accuracy metrics. Methods such as bootstrap resampling, repeated random splits, spatial cross-validation, or Monte Carlo experiments could be used to report the mean and 95% confidence interval of R^2, RMSE, MAE, and bias. This would make the performance assessment more robust and more convincing.

Reply： We sincerely thank the reviewer for this rigorous and constructive suggestion regarding uncertainty quantification.

We fully agree that reporting uncertainty ranges for model performance metrics is essential for robust assessment, as deep learning models are indeed subject to randomness in training, sample splitting, and parameter initialization.

In response to this comment, we conducted a bootstrap resampling analysis (1,000 iterations with replacement) to quantify the uncertainty of our main accuracy metrics. For the simulation dataset, we repeatedly resampled the test set and computed R^2, RMSE, MAE , Bias and RRMSE for all deep learning model configurations (Schemes A–D for GRU, and Scheme C for LSTM), reporting the mean values with their 95% confidence intervals. For the independent field validation, we applied the same bootstrap procedure to the combined 2022–2024 in situ measurements to capture the uncertainty arising from limited sample size and inter-annual variability. The detailed results have been incorporated into the revised manuscript as **Table A2 Bootstrap-derived uncertainty estimates (mean, standard deviation, and 95% confidence intervals) for deep learning model performance metrics** in **Appendix A**

Table A2 Bootstrap-derived uncertainty estimates (mean, standard deviation, and 95% confidence intervals) for deep learning model performance metrics.

Models	Phase	Scheme	Metric	Mean	Std	2.5% CI	97.5% CI
GRU	Simulation test	A(Mean Features)	R^2	0.453	0.003	0.447	0.460
			RMSE (t/ha)	1.070	3.257	1.064	1.077
			MAE (t/ha)	0.837	2.687	0.832	0.843
			Bias (t/ha)	-0.020	0.004	-0.029	-0.011
			RRMSE (%)	11.834	0.037	11.763	11.912
		B(Integral Features)	R^2	0.307	0.003	0.300	0.314
			RMSE (t/ha)	1.207	0.003	1.200	1.213
			MAE (t/ha)	0.958	0.003	0.953	0.963
			Bias (t/ha)	-0.070	0.005	-0.079	-0.061
			RRMSE (%)	13.379	0.038	13.304	13.451
		C(Full Features)	R^2	0.663	0.003	0.658	0.668
			RMSE (t/ha)	0.841	0.003	0.835	0.847
			MAE (t/ha)	0.645	0.002	0.640	0.649
			Bias (t/ha)	-0.019	0.003	-0.025	-0.013
			RRMSE (%)	9.323	0.035	9.257	9.392
		D(Selected Subset)	R^2	0.271	0.003	0.265	0.276
			RMSE (t/ha)	1.238	0.004	1.231	1.245
			MAE (t/ha)	0.987	0.003	0.980	0.993
			Bias (t/ha)	-0.004	0.005	-0.013	0.005
			RRMSE (%)	13.723	0.042	13.645	13.807
	Field Validation	A(Mean Features)	R^2	0.270	0.038	0.199	0.346
			RMSE (t/ha)	1.895	0.074	1.752	2.046
			MAE (t/ha)	1.458	0.058	1.348	1.578
			Bias (t/ha)	-0.432	0.085	-0.598	-0.264
			RRMSE (%)	21.475	0.835	19.921	23.173
		B(Integral Features)	R^2	0.495	0.058	0.384	0.607
			RMSE (t/ha)	2.195	0.136	1.950	2.465
			MAE (t/ha)	1.728	0.110	1.530	1.958
			Bias (t/ha)	-0.797	0.170	-1.135	-0.473

			RRMSE (%)	25.599	1.677	22.453	28.945
		C(Full Features)	R^2	0.689	0.021	0.649	0.728
			RMSE (t/ha)	1.210	0.028	1.155	1.264
			MAE (t/ha)	1.029	0.030	0.973	1.087
			Bias (t/ha)	-0.134	0.057	-0.244	-0.025
			RRMSE (%)	13.710	0.345	13.043	14.384
		D(Selected Subset)	R^2	0.308	0.037	0.239	0.384
			RMSE (t/ha)	1.808	0.063	1.689	1.933
			MAE (t/ha)	1.417	0.051	1.321	1.515
			Bias (t/ha)	0.758	0.077	0.603	0.918
			RRMSE (%)	20.490	0.840	18.899	22.176
LSTM	Simulation test	A(Mean Features)	R^2	0.565	0.003	0.559	0.571
			RMSE (t/ha)	0.955	0.003	0.949	0.962
			MAE (t/ha)	0.740	0.003	0.735	0.745
			Bias (t/ha)	-0.047	0.004	-0.054	-0.039
			RRMSE (%)	10.594	0.037	10.526	10.667
	Field Validation	A(Mean Features)	R^2	0.314	0.030	0.255	0.375
			RMSE (t/ha)	1.675	0.049	1.576	1.769
			MAE (t/ha)	1.352	0.045	1.263	1.437
			Bias (t/ha)	0.361	0.077	0.201	0.510
			RRMSE (%)	18.987	0.661	17.671	20.277

7.Explain how the model captured typhoon-related flood-induced yield losses in 2023.

The manuscript highlights the model's ability to capture yield losses caused by typhoon-related flooding in 2023. However, if WOFOST was mainly run under a water-limited mode, the simulated training samples would primarily represent water deficit or drought stress rather than yield losses caused by excessive soil moisture or waterlogging. The authors should explain which input signals allowed the model to detect flood-related yield reductions. Was the decline in LAI the main pathway? Were precipitation anomalies, soil moisture conditions, abrupt vegetation-index changes, or post-disaster canopy recovery also involved? If waterlogging processes were not explicitly represented in the simulations, this limitation should be clearly acknowledged in the discussion.

Reply： Thank you for this highly perceptive and insightful comment. We completely

agree that clarifying the physiological mechanism and key input signals that enabled the model to capture typhoon-induced flood losses in 2023 is essential for a accurate interpretation of our results.

In accordance with your suggestion, we have added a dedicated analysis in **Section 4.4 (*Simulation and Model Response to Extreme Precipitation Stress*)** to explain the underlying mechanism and explicitly acknowledge the model's inherent limitations:

4.4 Simulation and Model Response to Extreme Precipitation Stress

In this study, the WOFOST model was operated under the water-limited production mode, which is not equivalent to simulating only drought stress. The WOFOST soil water balance module is driven by actual daily precipitation and dynamically tracks changes in daily soil moisture content (van Diepen et al., 1989). Although the WOFOST model does not explicitly simulate the full range of physiological processes associated with anaerobic root conditions under waterlogging, this simplification does not prevent the model from capturing waterlogging-induced yield losses. When cumulative precipitation exceeds the soil drainage and runoff capacity, the model inherently triggers excess-water stress responses—suppressing root oxygen availability, limiting canopy transpiration and photosynthesis, and ultimately reducing LAI and biomass accumulation(H. et al., 1998; Liu et al., 2020). Therefore, our training dataset inherently includes yield reduction scenarios caused by excessive soil moisture, which originate directly from the 30-year historical meteorological records (1995–2024). Sentinel-2 satellite observations are capable of capturing such LAI anomaly signals, and the GRU model has learned, during the training phase, the general mapping relationship between depressed LAI trajectories and reduced final yields under various environmental stresses. When applied to real satellite imagery, the model does not explicitly detect flood events; instead, it infers yield losses through the observed LAI decline as a comprehensive stress signal.

8.Expand the range of WOFOST simulation scenarios.

Although the study generates a series of crop-model simulations, the simulation settings appear to be relatively simplified. For example, soil parameters are limited to four major loam types, and only four fixed sowing dates are considered. These assumptions may not fully capture the diversity of real-world field conditions. The authors should consider adding more combinations of sowing dates, soil parameters, water management, fertilization levels, and irrigation scenarios to improve the representativeness of the simulated training samples.

Reply： Thank you for this thoughtful suggestion regarding the expansion of simulation scenarios. We fully agree that the representativeness of simulated training samples is critical for the success of our field-label-free framework, and we appreciate the opportunity to clarify our simulation design rationale.

However, our current simulation settings, while appearing simplified, were deliberately designed to maximize the coverage of physiologically plausible growth trajectories through a full-factorial parameter combination strategy, rather than exhaustively enumerating all possible agronomic combinations, which would be computationally prohibitive and potentially introduce redundant samples.

Firstly, The four loam subclasses selected in our study (coarse-textured sandy loams, light-textured loams, intermediate-textured medium loams, and fine-textured heavy loams) represent the dominant soil texture classes across the Northeast China maize belt, covering over 85% of the cultivated area in the study region (Shi et al., 2004). While we acknowledge that soil heterogeneity exists at finer scales, the WOFOST model's hydrological module is primarily sensitive to soil texture-derived hydraulic parameters (wilting point, field capacity, saturation, and hydraulic conductivity).

Secondly, The four fixed sowing dates (April 20, April 30, May 10, and May 20) were selected to represent the typical range of actual sowing practices across Northeast China. Rather than arbitrarily increasing the number of sowing date scenarios, which would primarily shift the phenological timeline without fundamentally altering the shape of the growth trajectory, our full-factorial design combined these four dates with extensive crop parameter variations (192 crop parameter combinations), ensuring that the training dataset encompasses a wide range of growth rates, phenological durations, and stress responses.

Thirdly, the exhaustive factorial combination of multiple parameters (192 crop parameter combinations, 4 soils, 4 sowing dates, 60 station, 30 years, 3,686,400 valid samples) already ensures that the training dataset covers the vast majority of physiologically plausible growth trajectories under Northeast China's climatic and edaphic conditions. This scale is substantially larger than most previous surrogate modeling studies. Adding more soil subclasses or sowing dates would increase sample size but with diminishing returns in terms of physiological state space coverage.

Fourthly, as stated in Section 2.4.1, maize production in Northeast China is predominantly rainfed, and fertilizer application generally follows standardized regional practices that are not a limiting factor for growth. Therefore, operating the WOFOST model under water-limited mode with non-limiting nutrient assumptions is scientifically appropriate for the study region.

Ultimately, the strong performance of our model in independent field validation (R^2 = 0.69, RMSE = 1.21 t/ha) across multiple years provides empirical evidence that the current simulation settings are sufficiently representative to support accurate real-world yield estimation. We believe this outcome serves as the most compelling validation of our simulation design.

We have clarified this rationale in the revised manuscript by expanding the methodological justification in Section 2.4.2, emphasizing that our full-factorial design prioritizes physiological state space coverage over sheer scenario enumeration.

By systematically combining long-term meteorological records from 60 stations spanning 30 years, four major soil texture types, four representative sowing dates, and extensive crop parameter variations (192 crop parameter combinations), a maize *growth simulation dataset with pronounced spatiotemporal heterogeneity was constructed using the PCSE framework. Our full-factorial design prioritizes physiological state space coverage over exhaustive enumeration of all possible agronomic combinations, ensuring that the training dataset encompasses a wide range of growth rates, phenological durations, and stress responses under Northeast China's climatic and edaphic conditions. In total, 3,686,400 valid simulation samples were generated, and the detailed parameter combinations are summarized in Table 3. This dataset preserves complete time series trajectories of leaf area index (LAI) under diverse environmental stress conditions, together with the corresponding final grain yield, thereby providing a comprehensive crop growth simulation dataset for subsequent deep learning surrogate model training.*

9.Add validation at a finer administrative scale.

Validation based only on prefecture- or city-level yield statistics may be too coarse and could mask local spatial biases. I suggest collecting county-level maize yield statistics for the three northeastern provinces from 2019 to 2024, if available, and using them for additional validation. This would provide a more rigorous assessment of the spatial accuracy of the proposed yield product.

Reply： Thank you for your suggestion. We fully agree that city-level aggregation may obscure local spatial variations and fail to provide a rigorous assessment of spatial accuracy. In response to this comment, we have now acquired official county-level maize yield statistics for the three northeastern provinces for the period 2019–2024 and conducted an additional validation at the county scale. The detailed results have been incorporated into the revised manuscript in **Section 3.5.2 (Comparison with Existing Public Yield Products)**, along with the comparison against the existing 10-m benchmark dataset (Teng et al., 2026) , which provides a more rigorous assessment of the spatial accuracy of our proposed product.

The detailed results are presented in Section 3.5.2 of the revised manuscript, with quantitative county-level validation shown in Figure 14.

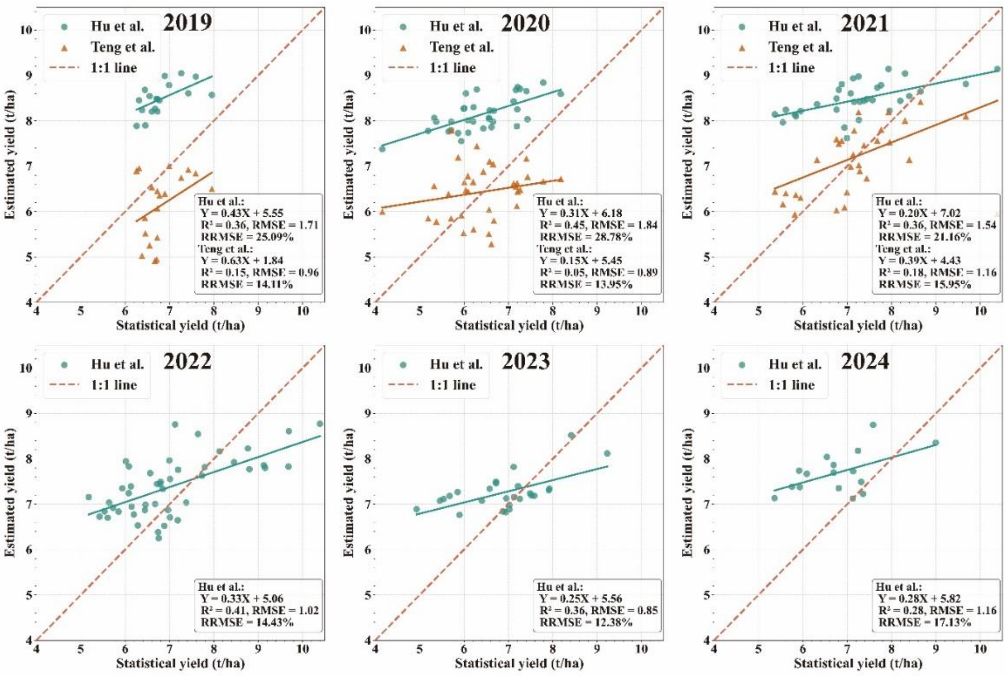


Figure 14 County-level comparison of maize yield estimates between our product and the existing 10-m benchmark dataset (Teng et al., 2026) against official statistical yearbook data in Northeast China (2019–2021).

We compared our product with the publicly available " A 10 m maize, rice and soybean yield dataset from 2016 to 2021 in Northeast China "(Teng et al., 2026) over the overlapping period (2019–2021). Both products were evaluated against official county-level statistical yearbook data (Figure 14).As shown in Figure 14, our product achieves consistently higher agreement with county-level statistics across all three years compared to the benchmark dataset. At the county-level, our product demonstrates robust performance across the full 2019–2024 period, with R^2 values ranging from 0.28 to 0.45 and RRMSE from 12.38% to 17.13%.

10.Discuss the simulation-to-reality gap more explicitly.

A key assumption of this study is that a deep-learning model trained on crop-model simulations can be successfully transferred to real remote-sensing conditions. Therefore, the gap between simulated samples and real-world fields is central to the reliability of the method. I suggest adding a dedicated discussion of this simulation-to-reality gap, including uncertainties related to crop-model parameters, management practices, remote-sensing LAI retrieval, meteorological inputs, and the absence or simplification of certain real-world stress processes such as flooding.

We sincerely thank the reviewer for raising this critical issue. We fully agree that the simulation-to-reality gap is central to the reliability of our proposed framework, as the fundamental premise of our approach—training a deep learning model entirely on simulated data and transferring it to real-world satellite observations—depends critically on the representativeness of the WOFOST simulations. In response to this comment, we have added a dedicated and comprehensive discussion to explicitly address this gap. Specifically, we have significantly extended Section 4.5 (Uncertainties and Limitations) to systematically enumerate the key discrepancies between the simulated and real domains.

4.5 Uncertainties and Limitations

Although this study achieved relatively high yield estimation accuracy, the model still carries certain uncertainties due to data source limitations and simplified assumptions. The gap between crop growth model simulations and reality, along with input data uncertainties, are the primary sources constraining the reliability of crop yield estimates.

The input layer of the WOFOST simulations, including both cultivar parameters and agronomic management settings, inevitably deviates from the complexity of real world conditions. In this study, several key physiological parameters governing phenological development were assigned based on literature values and the model's default settings, rather than calibrated against site-specific field trials for each cultivar. In terms of agronomic management, our framework assumes non-limiting nutrient conditions and discretizes the sowing window into only four fixed dates. In reality, smallholder farmers across Northeast China employ flexible sowing schedules and highly diverse management strategies (Chuanwei et al., 2024). These two types of input deviations may jointly cause the simulated LAI trajectories to diverge to some extent from actual growth curves in specific fields. Nevertheless, we argue that the full-factorial parameter combination strategy adopted in this study constructs an exhaustive training set that covers a broad physiological and management space. The consistency between estimated yields and statistical yearbook data (Figure 11), together with the robust multi-year independent field validation performance (Section 3.4), demonstrates that the simulated data generated under the current input settings are sufficiently reliable to support accurate yield estimation.

A systematic discrepancy also exists between the WOFOST-simulated LAI and the satellite- derived LAI. The simulated values represent a deterministic, noise-free "green leaf area index" at the point scale under idealized atmospheric conditions (H. et al., 1998). In contrast, the satellite-retrieved LAI is subject to multiple factors, including residual atmospheric correction errors, cloud-gap filling algorithms, and mixed-pixel effects. As noted by Lu et al. (2025), satellite-derived vegetation parameters (e.g., LAI) are not free from uncertainty, particularly in areas with high LAI or dense canopies, where saturation or underestimation may occur. Moreover, different data processing workflows, such as cloud masking and filtering algorithms, can introduce systematic errors. Consequently, input uncertainties can propagate through the prediction model when a model trained on clean simulated data is generalized to noisy satellite observations (Xie et al., 2025). However, the gating mechanism of the GRU provides *inherent robustness to such input noise by dynamically weighting the contribution of each time step (Cho, 2014). The stable performance of the model in independent validation across 2022–2024 (Table 7) further confirms that the proposed framework can effectively mitigate the adverse effects of uncertainties in satellite-derived LAI under the current data quality.*

Chen, Y., Wang, Y., Wu, C., Rosa Ferraz Jardim, A. M. da, Fang, M., Yao, L., Liu, G., Xu, Q., Chen, L., and Tang, X.: Drought-induced stress on rainfed and irrigated agriculture: Insights from multi-source satellite-derived ecological indicators, Agricultural Water Management, 307, 109249, https://doi.org/10.1016/j.agwat.2024.109249, 2025.

Cho, K.: Learning phrase representations using RNN encoder-decoder for statistical machine translation, arXiv preprint arXiv:1406.1078, 2014.

Chuanwei, Z., Gao, J., Liu, L., and Wu, S.: Simulating the effects of optimizing sowing date and variety shift on maize production at finer scale in northeast China under future climate, Journal of the Science of Food and Agriculture, 104, https://doi.org/10.1002/jsfa.13247, 2024.

van Diepen, C. a., Wolf, J., van Keulen, H., and Rappoldt, C.: WOFOST: A simulation model of crop production, Soil Use Manage., 5, 16–24, https://doi.org/10.1111/j.1475-2743.1989.tb00755.x, 1989.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., and Bargellini, P.: Sentinel-2: ESA's optical high-resolution mission for GMES operational services, Remote Sens. Environ., 120, 25–36, https://doi.org/10.1016/j.rse.2011.11.026, 2012.

Duo Q., Chengwei W., Xuemei B., Yanduo G., Qi S., Chen L., Kai T., and Yujie Z.: Characteristics and causes of extreme heavy rainfall in heilongjiang province during august 2023, yyqxxb, 35, 257–271, https://doi.org/10.11898/1001-7313.20240301, 2024.

H., B., Diepen, C. A., Rötter, R. P., and HH, van: WOFOST7.1: A user's guide, https://doi.org/10.13140/RG.2.1.1814.1282, 1998.

Hu, J., Du, X., Li, Q., Zhang, Y., Wang, H., Luo, J., Xu, J., Zhao, Y., Zhang, Z., Dong, Y., and Shen, Y.: NortheastChinaMaizeYield10m: A 10-m resolution maize yield dataset for northeast China (2019–2024) generated via a mechanistically interpretable, label-free framework, , https://doi.org/10.5281/zenodo.19547014, 2026.

Liu, K., Harrison, M., Shabala, S., Meinke, H., Ahmed, I., Zhang, Y., Tian, X., and Zhou, M.: The state of the art in modeling waterlogging impacts on plants: What do we know and what do we need to know, Earth's Future, https://doi.org/10.1029/2020EF001801, 2020.

Lu, J., Li, J., Fu, H., Zou, W., Kang, J., Yu, H., and Lin, X.: Estimation of rice yield using multi-source remote sensing data combined with crop growth model and deep learning algorithm, Agric. For. Meteorol., 370, 110600, https://doi.org/10.1016/j.agrformet.2025.110600, 2025.

Parker, G. G.: Tamm review: Leaf area index (LAI) is both a determinant and a consequence of important processes in vegetation canopies, Forest Ecology and Management, 477, 118496, https://doi.org/10.1016/j.foreco.2020.118496, 2020.

Teng, F., Wang, M., Shi, W., Pan, L., Guo, J., and Xiao, X.: A 10 m maize, rice and soybean yield dataset from 2016 to 2021 in northeast China, Sci Data, 13, 344, https://doi.org/10.1038/s41597-026-06719-0, 2026.

Xie, J., Zhang, D., Jin, N., Cheng, T., Zhao, G., Han, D., Niu, Z., and Li, W.: Coupling crop growth models and machine learning for scalable winter wheat yield estimation across major wheat regions in China, Agric. For. Meteorol., 372, 110687, https://doi.org/10.1016/j.agrformet.2025.110687, 2025.

Yin, J., Li, F., Li, M., Xia, R., Bao, X., Sun, J., and Liang, X.: The unique features in the 4 d widespread extreme rainfall event over north China in july 2023, Nat. Hazards Earth Syst. Sci., 25, 1719–1735, https://doi.org/10.5194/nhess-25-1719-2025, 2025.

Zhang, Y., Homma, K., Shi, L., Wang, Y., Qiao, H., and Zha, Y.: Improving crop modeling by simultaneously incorporating dynamic specific leaf area and leaf area index: A two-year experiment, Agricultural Systems, 227, 104357, https://doi.org/10.1016/j.agsy.2025.104357, 2025.