**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.*

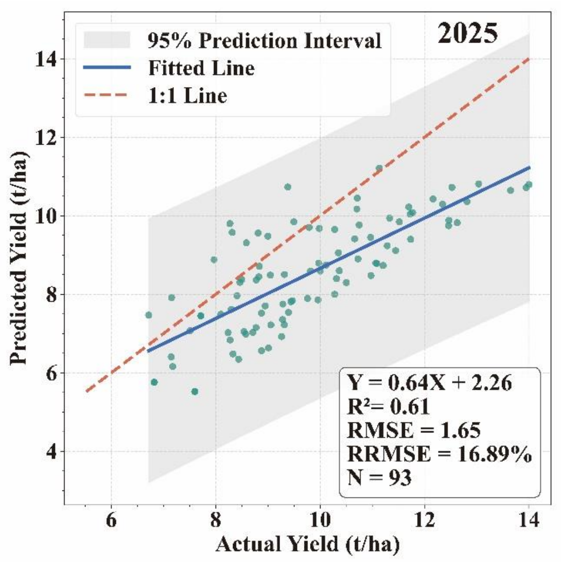


*Figure 9 Independent validation of the optimal GRU model (Scheme C) against in situ measurements in 2025.*

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 |

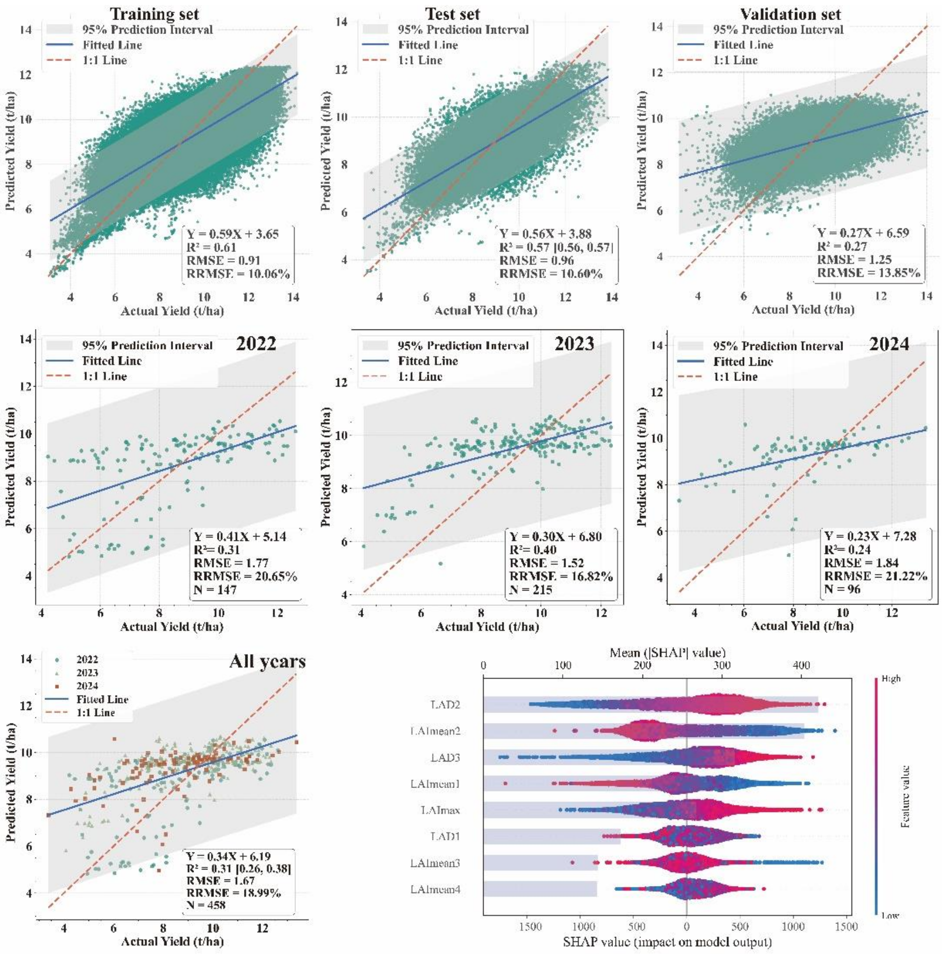


*Figure A2. Performance of LSTM models under Scheme C (Full Features) and independent validation using in situ yield measurements (2022–2024)*

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.*

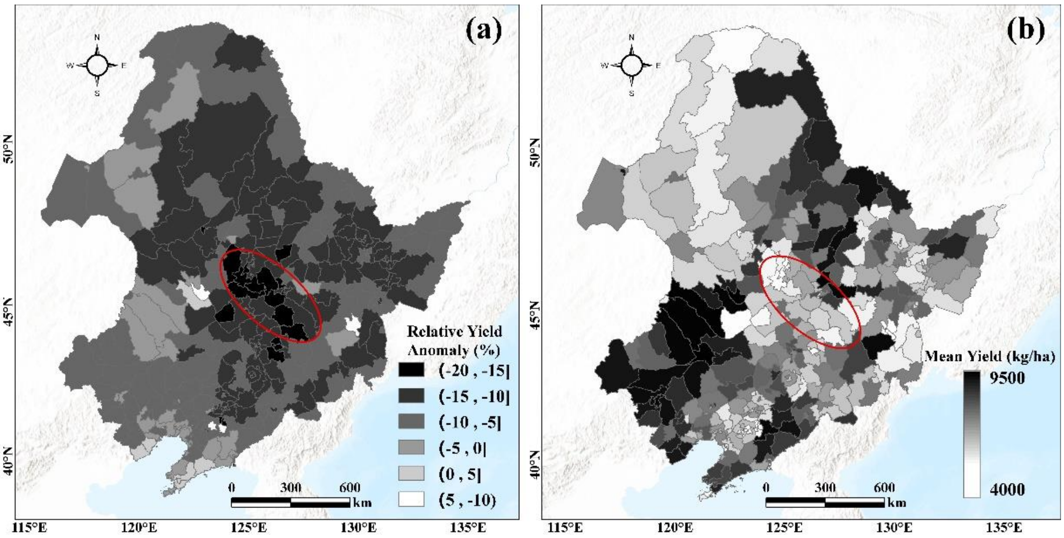


*Figure 12 County-level analysis of the 2023 flood-induced maize yield anomalies in Northeast China. (a) Spatial distribution of county-level yield anomalies in 2023 relative to the multi-year average (2019–2024). (b) Spatial distribution of estimated county-level maize yields in 2023. The most severe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