Author's response

We thank all the reviewers for their careful evaluation of our manuscript and for their thoughtful and constructive feedback, which has been invaluable in improving the quality and clarity of our work. We have carefully considered each comment and addressed them below. Please note that all line numbers referenced in our responses correspond to the clean version of the revised manuscript, not the tracked changes version.

Response to Reviewer #1 – *Geoscientific Model Development*
Manuscript EGUsphere-2024-4010
Authors: Jinhui Zheng, Le Yu *, Zhenrong Du, Liujun Xiao, and Xiaomeng Huang

Reviewer #1

Dear Reviewer Theodoros Mavromatis,

We sincerely thank you for your insightful comments and constructive suggestions, which have helped us improve the quality of our manuscript. Below, we provide detailed responses to each point. The original *reviewer comments are presented in italic*, while the authors' responses are provided in blue.

The line number is based on the clean version of the revised manuscript, not the track change version.

This is a significant contribution. Among others issues (noted on the attached manuscript that should be taken care of) a major one is related to the selection of the "extreme" years chosen in this study.

Response:

We sincerely thank the reviewer for the thoughtful and encouraging evaluation of our work, as well as for the valuable suggestions regarding the selection of extreme weather years. Your insightful comments have been instrumental in guiding us to improve the rigor and clarity of our analysis. In response to your recommendation, we have carefully re-evaluated our approach and made substantial revisions to the manuscript concerning the selection of representative years for extreme weather analysis. Drawing upon comprehensive survey data and official reports, we have now selected 2009, 2010, 2012, and 2018 as the representative extreme weather years used in this study. We believe these years more accurately and comprehensively reflect the diverse and significant impacts of extreme weather events on crop production in the study region. The updated rationale and supporting information can be found in lines 283–301 of the revised manuscript.

To briefly summarize:

According to the Ministry of Ecology and Environment of the People's Republic of China (www.mee.gov.cn), 2009 was marked by record-breaking high temperatures in the study area, with multiple locations exceeding historical maximums. In 2010, the frequency of meteorological disasters notably increased, and numerous extreme weather events were documented. In 2012, China experienced 38 severe rainfall events, including the devastating "7.21" event. The region also endured concurrent droughts and cold waves (Zhang et al., 2018; Zhao et al., 2019). In 2018, extreme low temperatures caused widespread frost damage, which had a pronounced impact on

agricultural productivity (China Meteorological Administration, www.cma.gov.cn).

We believe that this more cautious and evidence-based selection of extreme weather years provides a stronger foundation for evaluating the WOFOST-EW model under diverse and complex extreme weather conditions. These revisions have enhanced the scientific robustness, representativeness, and overall credibility of our study. Relevant references and background information have been added to support these changes. Once again, we sincerely appreciate your constructive feedback, which has greatly contributed to the improvement of our manuscript.

Lines 20-25: Report it as RRMSE as well.

Response:

Thank you very much for your careful review of the model performance evaluation metrics. We fully agree with your observation that reporting the Relative Root Mean Square Error (RRMSE) provides a standardized measure of relative error, which is highly valuable for assessing model performance and facilitating comparisons with other studies. Accordingly, we have calculated the RRMSE values as per your suggestion and added them to the abstract (lines 20–25). We have also included the corresponding explanations in the results section of the manuscript (lines 251–257, 259–282, and 296–301). We greatly appreciate your constructive feedback, which has helped improve the clarity and scientific rigor of our work.

Lines 20-25: see my comments in the text for these specific years.

Response:

Thank you for your valuable suggestion. As mentioned above, we have made substantial revisions to the study, including the re-selection of representative years for extreme weather events. Please refer to lines 283–301 in the revised manuscript for the updated content.

Lines 30-35: Which one? There are two in the reference list.

Response:

Thank you very much for pointing out this oversight. In response to your suggestion, we have revised the reference list to distinguish multiple publications by authors with the same surname published in the same year. Specifically, we have added “a,” “b,” etc., after the publication year to ensure that each reference is clearly identified and cited appropriately. Please refer to lines 33–34 in the revised manuscript for the updated citations.

Lines 35-45: Rephrase or delete. In my opinions, the previous sentences do not result in this one.

Response:

Thank you for your suggestion. As recommended, we have removed the sentence accordingly. Please refer to lines 45–47 in the revised manuscript.

Lines 65-70: Define this. What does it include?

Response:

Thank you very much for your valuable comment. We agree that the definition of "extreme climate" was not clearly articulated in the original manuscript, which may lead to confusion. In the revised manuscript, we have clarified this term explicitly (see lines 67–75).

In this study, "extreme climate" refers to abnormal temperature, precipitation, or drought events that occur during the entire growth season of winter wheat (from sowing to maturity). These events are characterized by their extremity, sudden onset, and damaging potential, which can significantly impact crop growth (Bai et al., 2024; Feng et al., 2019; Yu et al., 2025; Zheng and Zhang, 2025). Specifically, we used the following types of extreme climate indices as the basis for identification and simulation:

- Extreme temperature events: These include both heat and cold stress. Heat stress may cause premature senescence, poor grain filling, or even direct thermal damage, while cold stress can result in frost damage, delayed development, or yield loss. In our model, these impacts are quantified using indices such as high-temperature degree days (HDD) and low-temperature degree days (LDD) (Dong et al., 2023; Osman et al., 2020; Zhang et al., 2016; Zhang and Tao, 2019).

- Extreme precipitation events: These refer to unusually high-intensity rainfall over short periods or persistent abnormal precipitation patterns, which may lead to waterlogging, nutrient leaching, or increased disease pressure. We used indicators such as R95P (very wet days), R10 (number of days with precipitation >10 mm), and Rx1day (maximum 1-day precipitation) to characterize these events (Al-Sakkaf et al., 2024; Hong and Ying, 2018).

- Extreme drought conditions: These refer to prolonged periods of insufficient precipitation and/or high evapotranspiration that severely reduce soil moisture, causing water stress for crops. This can inhibit crop growth and photosynthesis and, in severe cases, lead to plant wilting or death. We used the Palmer Drought Severity Index (PDSI) and vapor pressure deficit (VPD) to represent drought conditions (Baydaroğlu et al., 2024; Kumar and Mahapatra, 2024; Oubaha et al., 2024; Peethani et al., 2024a; Pei et al., 2024a; SM et al., 2025; Yang et al., 2024; Zhang et al., 2025).

These events are defined as "extreme" because their intensity, duration, or frequency significantly exceed the normal historical range, with clear evidence of their adverse effects on wheat physiological processes and final yield. The WOFOST-EW model introduced in this study is designed to better simulate and assess the impact of these specific types of extreme climate conditions on wheat production. To eliminate any ambiguity, we have included this definition of

"extreme climate" in lines 67–75 of the revised manuscript.

Lines 75-85: Elaborate on this. Mention a few dynamic changes in crop growth which are overlooked.

Response:

Thank you very much for your constructive suggestion. We have incorporated specific revisions in lines 80–87 of the revised manuscript to clarify these dynamic processes.

Specifically, in most previous studies, the outputs of crop models (such as biomass, leaf area index, or final yield) were often directly used as input variables for machine learning models. However, little attention has been paid to how extreme weather events nonlinearly disturb crop physiological and developmental processes at different growth stages. This form of static coupling overlooks the temporal sensitivity and continuity of extreme climate impacts. Phenological responses, for example, are often not fully considered—extreme heat or drought may lead to earlier heading or maturity in wheat (Hou et al., 2024; Liu et al., 2023), which can significantly affect dry matter accumulation. These nonlinear dynamics are frequently neglected in conventional machine learning frameworks.

Recognizing the omission of these critical dynamic responses, our study aims to enhance the phenology module of the WOFOST-EW model by integrating extreme weather indices with deep learning algorithms. This approach enables the model to more accurately and robustly capture the complex, stage-specific effects of extreme weather events on crop growth processes, thereby improving yield prediction accuracy under extreme climatic conditions.

Lines 75-85: Mention a few appropriate references for this statement.

Response:

Thank you for your suggestion. We have added relevant references to support this statement. Please refer to lines 80–87 in the revised manuscript.

Table 1: Move this Table in the Supplementary material. Presenting growing season temperature and precipitation for each station, would be more relevant.

Response:

Thank you for your valuable suggestion. We have moved the table to the Supplementary Materials as Table S1 and made the corresponding adjustments in the main text. Additionally, we have included information on the temperature and precipitation during the winter wheat growing season for each station. Please refer to Table S1 in the Supplementary Materials.

Lines 105-110: Mention them.

Response:

Thank you for your suggestion. We have added details about the main elements included in the weather dataset. Please refer to lines 105–109 in the revised manuscript.

Lines 105-110: How about available water capacity for each layer?

Response:

The soil data from the ISRIC global database include Available Water Capacity (mm/m). We have explicitly added this important information to the revised manuscript to provide readers with a more comprehensive understanding of the soil data details. Please refer to lines 118–121 in the revised manuscript.

Lines 120-125: In years 2008 and 2018.

Response:

Thank you for your suggestion. We have made the corresponding additions in the revised manuscript. Please refer to lines 130–132.

Lines 140-145: Why these two indices were preferred over the others included in the CHM_Drought database?

Response:

Thank you for your valuable question. We selected the Palmer Drought Severity Index (PDSI) and Vapor Pressure Deficit (VPD) from the CHM_Drought database for several key reasons:

First, PDSI is one of the most widely used drought indices (Oubaha et al., 2024; Yang et al., 2024; Zhang et al., 2025). It accounts for antecedent precipitation and water supply-demand balance, providing clear physical meaning and is particularly suitable for assessing agricultural drought. In this study, we focus on the winter wheat growing season, and PDSI effectively captures drought processes at this timescale. Numerous studies have demonstrated its significant correlation with crop yield (Baydaroğlu et al., 2024; Kumar and Mahapatra, 2024; Peethani et al., 2024a; Pei et al., 2024a; SM et al., 2025).

Second, VPD is a key variable measuring atmospheric dryness, directly affecting crop transpiration and water stress. Studies have reported a steady global increase in VPD from 2010 to 2019, severely hindering agricultural production (Koehler et al., 2023; Nesmith and Ritchie, 1992). Under extreme heat and low humidity, elevated VPD exacerbates crop transpiration and water loss, posing a direct threat to yields. Thus, VPD can partly reflect the stress intensity of short-term extreme heat and

drought events (Yu et al., 2024).

Although the CHM_Drought database provides various indices, our literature review and preliminary analyses show that these two indices exhibit stronger applicability and higher historical validation reliability for drought monitoring and assessment in the North China Plain (Li et al., 2024; Luan et al., 2024; Wu et al., 2024). They better correspond with observed agricultural drought events and yield losses in the region.

We have further elaborated on the rationale for selecting these indices in lines 146–158 of the revised manuscript. We appreciate your insightful comment, which helped us clarify the methodological choices in our study.

Lines 190-195: Are these extreme indices are estimated every day and then be used an input to LSTM? How about PDSI and VPD? On which temporal basis are estimated? Why not the PDSI over the ScPDSI? Why R95P, R10 and Rxwday are estimated on annual and not on growing season basis? Elaborate on these.

Response:

Thank you very much for your detailed question regarding the processing of extreme weather indices. First, we would like to clarify that all extreme indices used in this study—including PDSI, VPD, R95P, R10, Rx1day, and others mentioned later—were calculated over the entire growing period of winter wheat. This approach ensures that the extreme weather events we analyze are closely linked to the actual growth and developmental stages of the crop, allowing for a more accurate assessment of their impacts on yield. Below is a detailed explanation based on the manuscript content:

> *Are these extreme indices are estimated every day and then be used an input to LSTM? How about PDSI and VPD? On which temporal basis are estimated?*

Response:

In our study, high-temperature degree days (HDD) and low-temperature degree days (LDD) are cumulative indices based on wheat growth stages. Specifically, the high-temperature thresholds were set at 25°C from sowing to heading and 30°C from heading to maturity; the low-temperature thresholds were −5.7°C and −2°C for the corresponding periods (Farooq et al., 2011; Liu et al., 2013; Porter and Gawith, 1999). These indices are calculated daily by summing temperature deviations and directly reflect the sustained impact of extreme temperatures during critical phenological stages. Their timing strictly corresponds to the wheat growth cycle, making them suitable as input features for the LSTM to characterize stage-specific climate stress intensity. Although extreme precipitation indices such as R95P, R10, and Rx1day are often reported on an annual scale in the literature, we recalculated them based on the wheat growing season to more accurately capture precipitation extremes encountered during crop growth. The Palmer Drought Severity Index (PDSI) reflects long-term soil moisture conditions by integrating precipitation, temperature, and potential

evapotranspiration, and is typically calculated on a monthly or seasonal basis (Zhang and Miao, 2024). In this study, PDSI and vapor pressure deficit (VPD) were averaged or accumulated over the growing season to represent the sustained effect of soil moisture deficits. We acknowledge that the original manuscript's descriptions were incomplete; thus, we have revised the definitions in Table 1 for greater accuracy and supplemented the explanations in lines 135–162 of the revised manuscript to improve clarity.

Why not the PDSI over the ScPDSI?

Response:

The Self-Calibrating Palmer Drought Severity Index (ScPDSI) modifies the empirical constants used in the original PDSI calculation by dynamically adjusting them to local climate conditions, allowing it to automatically calibrate drought behavior at any location (Dai, 2011; Zhang and Miao, 2024). However, studies have shown that in most regions of China, PDSI exhibits a stronger correlation with normalized difference vegetation index (NDVI) anomalies, simulated soil moisture anomalies (SMA), and the land water storage deficit index (WSDI) compared to ScPDSI (Zhong et al., 2019). This suggests that PDSI is more representative for characterizing agricultural drought impacts on crops in the Chinese context. One main reason is that ScPDSI tends to capture meteorological droughts as less severe than PDSI, which may be attributed to two factors: 1) modifications in ScPDSI reduce sensitivity to different potential evapotranspiration (PET) estimation methods, leading to lower responsiveness under wet or dry conditions (van der Schrier et al., 2011); 2) adjustments to the self-calibrating persistence factors and climate characteristic parameter may increase ScPDSI's sensitivity to dataset-specific features, sometimes weakening its drought detection performance in certain regions. For example, Liu et al. (2016) confirmed this in the Yellow River Basin of northern China. Although the self-calibrating procedure improves spatial consistency and controls extreme event frequency (Dai, 2011b; Trenberth et al., 2014), considering our study's specific objective to evaluate WOFOST-EW model performance under extreme conditions, we argue that PDSI sufficiently and effectively reflects drought conditions impacting crop growth within the current data and regional context. Moreover, PDSI has been widely validated in numerous crop-related studies (Islam et al.; Peethani et al., 2024b; Pei et al., 2024b; Yan et al., 2016). To clarify this choice further, we added a detailed explanation in lines 146–162 of the revised manuscript.

Why R95P, R10 and Rxwday are estimated on annual and not on growing season basis? Elaborate on these.

Response:

Thank you very much for pointing out the potential ambiguity in our description. We would like to clarify that R95P, R10, and Rx1day in our study were recalculated specifically for the winter wheat growing season to more directly capture the extreme precipitation events encountered during the crop growth period. We have revised the definitions in Table 1 as well as the related content in lines 135–162 to avoid any confusion for readers. Once again, we sincerely appreciate your constructive

comments and careful review.

Lines 200-205: between measured and estimated yield?

Response:

Thank you for pointing out this issue. The statement refers to the approach where the parameters corresponding to the minimum root mean square error (RMSE) between observed and simulated yield, as well as between observed and simulated phenological stages, are considered optimal. This is a standard method for assessing model fit during calibration and optimization (Chen and Tao, 2020; Zheng and Zhang, 2023). We have revised this sentence to improve clarity and accuracy. Please refer to lines 214–225 in the revised manuscript.

Lines 200-205: Which parameters of the Table S2 were calibrated and for which target (anthesis, maturity and yield)?

Response:

We sincerely thank the reviewer for the careful review and valuable comments. We have added supplementary explanations and clarifications in the revised manuscript; please refer to lines 214–225 in the main text as well as Text S1, S2 and Tables S3, S4 in the supplementary materials. Specifically, we included the sensitivity analysis results of the WOFOST model (Table S3) and identified the parameters to be adjusted based on this analysis (Table S4). We appreciate your important questions, which helped us better clarify the rationale and optimization process of the model parameter settings.

Lines 210-215: Estimate and show the respective results for MAE (mean absolute error). How about the results for r2?

Response:

We fully agree with your viewpoint that Mean Absolute Error (MAE) and the Coefficient of Determination (R^2) are important metrics for evaluating model prediction accuracy and explanatory power. Accordingly, we have calculated MAE and R^2 as per your suggestion and have detailed these evaluation results in lines 230–257 of the revised manuscript.

Lines 245-250: Express it and in the form or relative RMSE (within parentheses).

Response:

Thank you very much for your valuable suggestion. Following your advice, we have calculated the RRMSE and included it in parentheses in lines 259–273 of the revised manuscript to more clearly demonstrate the model's performance.

Figure 5: Which statistic is this? Has this been described in section 2.3.6?

Response:

Thank you very much for your question. The statistic shown in Figure 5c is the simulation bias, defined as the difference between simulated and observed yields (simulated value minus observed value). This metric is used to assess the model's systematic overestimation or underestimation of yield across different years. We have added a clear definition of this statistic in Section 2.3.5 of the revised manuscript, along with its corresponding formula (Equation 16).

Lines 255-260: Is this for the calibration or validation period?

Response:

Thank you very much for your question. The statistical results shown in Figure 5c indeed correspond to the validation period. To avoid any confusion, we have revised the figure caption in the updated manuscript to explicitly indicate the time period represented. Please refer to the updated Figure 5 in the revised manuscript.

Lines 260-265: This statistic should be defined in 2.3.6 section. Express it also in its relative version (relative MAD).

Response:

Thank you very much for your valuable suggestion. We have explicitly added the definition of the MAE statistic in Section 2.3.5 of the revised manuscript. Additionally, following your recommendation, we have further calculated and reported the Mean Relative Error (MRE) to more intuitively represent the relative proportion of simulation errors to observed values. This metric helps to better understand the model performance across different counties and years. The related results have been added in lines 270–275 of the revised manuscript.

Lines 265-270: Which statistic is this? Has this been described in section 2.3.6?

Response:

Thank you very much for your valuable suggestion. We have explicitly added the definition of the Mean Absolute Error (MAE) in Section 2.3.5 of the revised manuscript. Additionally, following your advice, we further calculated and reported the Mean Relative Error (MRE) to provide a more intuitive representation of the simulation errors relative to the observed values. This metric helps better understand the model's performance across different counties and years. The related results have been added in lines 269–273 of the revised manuscript.

Figure 7: 2001?

Response:

Thank you for pointing this out. The error has been corrected, and the entire manuscript has been carefully reviewed. Please refer to Figure 7 in the revised manuscript.

Lines 275-280: I am not sure that the selected years are the best options. The observed yield of 2018 (see my red rectangle) is one of the highest in the study period. That means that wheat in reality recovered from the extreme weather noted in 2018 (Table S1) at many stations as can be seen from Fig. 8c, d. 2008 yield is also one of the highest in the study period. Maybe different years should be selected.

Response:

We sincerely appreciate the reviewer's careful evaluation and valuable suggestion regarding the selection of extreme weather years. Based on your insightful comments, we have re-examined the choice of representative years to better reflect the actual impacts of extreme weather on wheat yields in the study region. After a thorough review of official reports and observational data, we have revised our selection to include 2009, 2010, 2012, and 2018 as the key years representing extreme weather events (see lines 283–301 and Figs. 9 and 10 in the revised manuscript).

While it is true that the observed yield in 2018 was relatively high, this year was characterized by severe frost damage and other extreme low-temperature events documented by official sources (China Meteorological Administration). Our analysis also considers the spatial and temporal variability of impacts, as shown in Fig. 8, which captures differing regional responses to the extreme conditions (Ministry of Ecology and Environment of the People's Republic of China). The years 2009, 2010, and 2012 correspond to well-documented episodes of record-breaking heat, increased meteorological disasters, and extreme rainfall events, respectively (Zhao et al., 2019; Zheng et al., 2018).

We believe that this selection better balances the representation of diverse extreme weather types and their influence on crop production across the region. Detailed justification and supporting data have been added in lines 283–301 of the revised manuscript. We thank the reviewer again for the constructive advice, which has strengthened the rigor and clarity of our study.

Lines 280-285: No such counties in Table S1.

Response:

Thank you very much for your careful observation. The previous wording was indeed inaccurate. Shanxi and Hebei refer to provinces rather than county names. We have revised the relevant statements (lines 283–301 in the revised manuscript) to ensure clearer and more accurate expression, avoiding any potential confusion. We sincerely appreciate your meticulous review and valuable comments.

Figure 8: yield

Response:

Thank you for your suggestion. We have revised the figure caption accordingly. Please refer to Figure 9 in the revised manuscript.

Lines 310-315: Maybe estimating R95P, R10 and Rxwday on growing season basis rather on annual could help.

Response:

Thank you very much for your valuable comment. We fully understand your concern and would like to clarify that all extreme climate indices used in this study—including R95P, R10mm, and Rx1day—are calculated specifically for the winter wheat growing season (from sowing to maturity), rather than on an annual basis. This approach ensures that the identified extreme events are closely aligned with the actual crop growth period, allowing for a more accurate assessment of their potential impact on yield. We have added detailed explanations in lines 135–162 and updated Table 1 in the revised manuscript to prevent any possible misunderstanding.

Lines 320-325: I am not sure I understand this statement. Elaborate and make it more clearly. An appropriate reference would help.

Response:

Thank you for pointing out this issue. What we intended to convey is that the increasing frequency of extreme weather events leads to greater variability and unpredictability in meteorological observations such as temperature and precipitation. This variability complicates the derivation of stable and representative input parameters for crop models—such as thermal time and stress thresholds—thereby introducing uncertainty into model simulations (Gao et al., 2020; Gao et al., 2021). Such instability may cause deviations in model outputs and ultimately affect the accuracy of crop growth predictions. We have revised the original unclear statements and added relevant references accordingly. Please refer to lines 370–383 in the revised manuscript.

Lines 325-330: see my previous comment.

Response:

Thank you for raising this point. Here, we intended to express that in the North China Plain, frequent extreme high and low temperatures disrupt the consistency of daily weather inputs used in the model (Gu et al., 2024). This inconsistency affects the reliability of key model parameters, such as effective temperature accumulation and phenological thresholds, ultimately reducing the accuracy of crop growth and yield simulations under these extreme conditions (Bai et al., 2024). We have revised the previously unclear statements and supplemented relevant references accordingly. Please refer to lines 375–380 in the revised manuscript.

Lines 325-330: Rephrase.

Response:

Thank you for your suggestion. This sentence has been appropriately revised. Please refer to line 385 in the revised manuscript.

Lines 340-345: This is an overstatement. Only anthesis, maturity and final yield were checked. Rephrase.

Response:

Thank you for your suggestion. This sentence has been appropriately revised. Please refer to lines 390–395 in the revised manuscript.

Lines 340-345: No comparison on long-term trends between WOFOST and WOFOST-EW was made. Rephrase.

Response:

Thank you for your suggestion. This sentence has been appropriately revised. Please see lines 390–395 in the revised manuscript.

Lines 345-350: see a previous comment for these years.

Response:

Thank you very much for your valuable suggestion. In response to your comment, we have carefully re-examined the criteria for selecting extreme years and made substantial revisions accordingly. The related content has been thoroughly updated in lines 284–301 of the revised manuscript, along with corresponding updates to the figures and supplementary materials. Please refer to our detailed response provided earlier.

Lines 350-355: Does the LSTM algorithm estimates one F(EW) value for each site? That means that this value is site- and study period- specific.

Response:

Thank you very much for raising this important point. We would like to provide further clarification as follows:

In this study, the LSTM model is trained separately for each county to fully capture the spatial heterogeneity of extreme weather impacts on crop yield. The model inputs include multiple extreme climate indices during the winter wheat growing season (e.g., HDD, LDD, R10mm, PDSI, etc.), and the output is the extreme weather impact factor, F(EW), calculated for each year and county. This factor quantifies the integrated effect of extreme climate conditions on that year's yield, making F(EW) a spatiotemporally dynamic output variable.

While the WOFOST-EW framework can, in principle, be extended to other crops and regions, there are some inherent limitations that we have discussed in lines 399–438 of the revised manuscript. First, although WOFOST is a generic crop growth model, its parameters and certain modules are more specifically tailored to particular crop types such as cereals. For structurally different crops—such as root/tuber crops, oilseed crops, or perennials—adjustments to WOFOST's internal parameters would be necessary. Second, crop growth characteristics vary across regions due to differences in local climate, soil conditions, and management practices. Even if the WOFOST model structure is applicable, detailed parameter localization and calibration would be required when applying it to new crop types and target areas.

Furthermore, crop growth is influenced not only by climatic factors but also strongly affected by soil fertility, pest and disease pressures, irrigation, fertilization, and other complex environmental and management variables. Currently, the extreme weather functions within WOFOST-EW primarily emphasize meteorological factors. Effectively integrating non-meteorological extreme stresses—such as sudden pest outbreaks or severe nutrient deficiencies—and their interactions with extreme weather remains an important direction for future research.

We have clarified the model structure and the interpretation of F(EW) in lines 164–178 of the revised manuscript, and have emphasized the model's regional adaptability and limitations accordingly.

Lines 360-365: Did the LSTM algorithm only aimed at minimizing the error between observed and estimated yield?

Response:

Thank you for your question. The role of the LSTM algorithm in this study is not simply to directly minimize the error between observed and simulated yields. Specifically, the core objective of the LSTM is to learn and estimate a spatiotemporally dynamic extreme weather function F(EW). This F(EW) variable is designed to capture how various extreme weather indicators—such as extreme temperatures, precipitation, and drought—complexly affect crop growth processes.

The LSTM learns the nonlinear dynamic relationships between extreme weather events and crop physiological responses, thereby generating a correction factor that adjusts the internal growth processes within the WOFOST model in real time. In other words, the training goal of the LSTM can be more accurately described as optimizing the estimation of F(EW), enabling the coupled WOFOST-EW model system to simulate crop growth dynamics and final outcomes—including phenological stages (e.g., heading, maturity) and final yield—more accurately under extreme climate conditions.

Thus, by learning the modulation effects of extreme weather on crop growth, the LSTM indirectly helps minimize the discrepancy between observed and simulated yields, while more finely capturing crop responses under extreme conditions.

We have elaborated on the specific role of the LSTM algorithm, its integration with the WOFOST model, and its training objective in the revised manuscript (see lines 164–178 and 209–211). Thank you again for your insightful question, which prompted us to clarify the internal logic of our model more clearly.

Lines 390-395: Report it as RRMSE as well.

Response:

Thank you for your suggestion. We have added the RRMSE values accordingly. Please refer to lines 440–447 in the revised manuscript.

Response to Reviewer #2 – *Geoscientific Model Development*
Manuscript EGUsphere-2024-4010
Authors: Jinhui Zheng, Le Yu *, Zhenrong Du, Liujun Xiao, and Xiaomeng Huang

Reviewer #2

Dear Reviewer,

We sincerely thank you for your insightful comments and constructive suggestions, which have helped us improve the quality of our manuscript. Below, we provide detailed responses to each point. The original *reviewer comments are presented in italic*, while the authors' responses are provided in blue.

The line number is based on the clean version of the revised manuscript, not the track change version.

The paper presents WOFOST-EW, an enhanced version of the WOFOST crop model that employs an LSTM to integrate extreme weather indices, aiming to improve the phenology component of WOFOST. The model is validated using data from the North China Plain, focusing on the prediction of two phenological stages (heading and maturity) and yield, particularly during extreme weather years. While the approach of integrating machine learning into a process-based crop growth model is interesting, the implementation and evaluation raise several significant concerns, outlined below.

Response:

Thank you very much for your valuable overall evaluation of our work. We are glad to know that you recognize our efforts. At the same time, your comments are extremely important to us. We have carefully addressed each of the issues you raised to ensure the quality of the manuscript and the rigor of the research.

Use of Extreme Weather Indices in phenology prediction

● *The authors propose that extreme weather events, such as excessive rainfall or temperature anomalies, can improve the prediction of phenological stages. However, it is commonly understood that phenological development in cereal crops is primarily driven by temperature (thermal time), photoperiod, and vernalization. To strengthen their argument, the authors should provide a clear rationale explaining why and how extreme weather events would directly affect phenological stages.*

Response:

Thank you very much for your insightful and important question regarding how extreme weather indices influence phenology prediction in this study. You accurately pointed out the widely accepted

understanding that the phenological development of cereal crops is primarily driven by temperature (thermal time), photoperiod, and vernalization. We fully agree with your view, which reflects the commonly adopted physiological mechanisms in current crop models. In response to the approach we proposed in the manuscript—introducing extreme weather indices for phenology prediction—we would like to provide a more detailed explanation and clarification here.

In actual agricultural production, although thermal time is the main factor controlling phenological changes, an increasing number of empirical studies have shown that extreme weather events (such as extreme heat, cold, drought, and heavy rainfall) can significantly disturb the physiological processes of crops, resulting in either advancement or delay in phenology (Chachar et al., 2016; Chowdhury et al., 2021; Harrison, 2021; Ihsan et al., 2016; Li et al., 2020; Liu et al., 2023; McMaster and Wilhelm, 2003; Porter and Gawith, 1999; Xiao et al., 2021). Below are the reasons and mechanisms through which extreme weather events directly or indirectly affect phenological stages:

- Effects of Extreme Temperature on Wheat Phenology

  Numerous crop modeling studies and empirical analyses have shown that high temperatures during the growing season generally shorten the duration of crop development (Asseng et al., 2015; Sadras and Monzon, 2006; Tao and Zhang, 2013; Zahra et al., 2021). However, the effects of warming vary across phenological stages. For example, the study by Li et al. (2020) found that warming significantly shortened the period from sowing to flowering, while the period from flowering to maturity was relatively less affected. Generally, the pre-flowering stage is more influenced by minimum temperatures, while the post-flowering stage is more affected by maximum temperatures (Porter and Gawith, 1999). Moreover, wheat development is more sensitive to temperature changes during the pre-flowering phase, which coincides with the period when temperature increases are most pronounced (Tao et al., 2017a; Tao et al., 2017b). During this stage, moderate warming can accelerate thermal accumulation, promote earlier heading and flowering, and change the rhythm of the entire phenological process. Some studies also suggest that moderate warming can enhance enzyme activities related to photosynthesis and promote leaf area expansion (Chen et al., 2014), thereby increasing photosynthetic capacity and stimulating LAI growth and chlorophyll accumulation, which could lead to earlier flowering (Li et al., 2020). However, after flowering, high temperatures tend to accelerate leaf senescence, reduce chlorophyll content and photosynthesis rates, and ultimately shorten the grain-filling period, resulting in earlier maturity. Notably, recent extreme high-temperature events have caused significant disruptions to wheat growth. For instance, Liu et al. (2023) pointed out that the extreme heat in the summer of 2023 in China exceeded the upper threshold for optimal crop growth (Harrison, 2021), directly causing early maturity of wheat. These nonlinear and stage-specific temperature response mechanisms suggest that under extreme weather conditions, relying solely on thermal time to simulate wheat phenology

may be insufficient.

- Effects of Extreme Cold on Wheat Phenology

  Extreme low temperatures, particularly frost events below crop tolerance thresholds, can significantly disrupt wheat phenological development. Frost may directly damage critical growth organs, such as young spikes and reproductive tissues, interrupting development or causing notable delays (Fuller et al., 2007). During early vegetative growth, frost can dramatically reduce seedling survival and lead to leaf burn and necrosis, commonly manifested as leaf tip scorching (Shroyer et al., 1995). More importantly, rapid temperature drops often have more damaging effects than gradual declines, as crops lack sufficient time to acclimate to sudden cold conditions (Al Issawi et al., 2013; Li et al., 2014) In such cases, even if the temperature does not reach lethal levels, latent damage to organ formation may occur, hindering subsequent development. Studies have shown that low-temperature stress is one of the major limiting factors for global wheat production. Persistent low temperatures suppress metabolic activity and slow tissue division and elongation, thereby significantly prolonging phenological stages—particularly from jointing to heading (Xiao et al., 2021). Moreover, if frost or cold waves occur during stem elongation or spikelet initiation stages, they may lead to developmental disorders, delayed heading, or even partial sterility, severely affecting final yield.

- Effects of Drought/Water Stress on Wheat Phenology

  Although temperature is the dominant environmental driver of wheat development, drought or water stress can also significantly affect wheat phenology, especially in drought-prone regions (McMaster and Wilhelm, 2003). In recent years, increasing attention has been paid to the potential disruptions caused by drought stress on crop developmental rhythms, and drought is increasingly regarded as a crucial environmental variable affecting crop phenolog y (McMaster and Smika, 1988). The effects of drought stress on crop development depend on the timing, developmental stage, and severity of the stress. When wheat encounters moderate to severe water stress during critical developmental phases, it triggers a series of morphological, physiological, biochemical, and molecular stress responses, which can alter its normal growth rhythm (Chachar et al., 2016; Ihsan et al., 2016). These responses include decreased photosynthesis, stomatal closure, abnormal hormone secretion, and impaired reproductive development, which may ultimately accelerate or disrupt developmental progress. To adapt to water stress, wheat has evolved various drought-resistance strategies, including drought escape (accelerating the life cycle to avoid drought), drought avoidance (e.g., regulating stomatal behavior to reduce water loss), and drought tolerance (maintaining cellular function under stress) (Nyaupane et al., 2024). Although these strategies help ensure survival and a degree of yield stability, they often result in shortened developmental cycles and lead to earlier phenology (Chachar et al., 2016; Chowdhury et al., 2021; Ihsan et al., 2016; McMaster and Wilhelm, 2003).

- Effects of Extreme Rainfall/Flooding on Wheat Phenology

  Extreme rainfall events often cause field flooding or waterlogging, which can significantly hinder normal crop development (Liu et al., 2020). Waterlogging leads to hypoxia in the soil, thereby reducing oxygen supply to plant roots (Colmer and Greenway, 2011; Kotula et al., 2015), inhibiting root elongation and function, and limiting nutrient uptake and transport to aboveground organs (Colmer and Voesenek, 2009). In severe cases, it may even result in partial root death (Herzog et al., 2016). Additionally, excessive rainfall can reduce root elongation rates (Ebrahimi-Mollabashi et al., 2019) and leach essential nutrients from the soil (Kaur et al., 2020), causing nutrient stress. Even in the absence of visible nutrient deficiencies, waterlogging affects the roots' ability to absorb and transport water, leading to stomatal closure, decreased internal leaf CO_2 concentration, and reduced photosynthesis and crop growth (Jitsuyama, 2017). Furthermore, water accumulation increases the risk of pathogen proliferation, potentially triggering root rot, leaf diseases, and increasing lodging risk (Nguyen et al., 2016). Although most existing studies focus on the effects of waterlogging on crop growth and yield, some scholars have called for greater attention to its impact on crop phenological development (Nóia Júnior et al., 2023). Empirical studies have shown that flooding during tillering and jointing stages of wheat can significantly inhibit chlorophyll synthesis and photosynthesis, obstructing early growth and potentially delaying subsequent phenological stages such as jointing and heading (Dickin and Wright, 2008; Wu et al., 2015).

Our WOFOST-EW model introduces an LSTM algorithm and uses multiple extreme weather indices (such as HDD, LDD, PDSI, VPD, and R95P) as inputs to learn the dynamic modulation effects of these extreme events on crop phenological processes. The LSTM can capture the nonlinear relationships between extreme events and phenological phase changes, generating a temporally dynamic variable F(EW). This F(EW) acts as a correction factor for the phenology module within the WOFOST model, allowing the model to further refine its simulation of acceleration, delay, or interruption of phenological phases induced by extreme weather, on top of the standard thermal time, photoperiod, and vernalization mechanisms. This enhancement improves the robustness and accuracy of phenology predictions under extreme climate conditions.

We have clarified and elaborated on this theoretical foundation in lines 320–368 of the revised manuscript, supported by relevant literature. Once again, thank you for your profound question, which has prompted a more complete explanation of the scientific principles underlying our model.

● *It appears that these extreme weather indicators are not incorporated into other model components (e.g., growth, biomass accumulation, or stress responses), but are instead used solely to modulate phenology. This approach attributes all extreme weather effects to phenology alone, which risks conflating physiological processes and may compromise the model's interpretability.*

Response:

Thank you very much for your thoughtful review and constructive suggestions. In response to your concerns, we would like to provide the following detailed explanation.

> *It appears that these extreme weather indicators are not incorporated into other model components (e.g., growth, biomass accumulation, or stress responses), but are instead used solely to modulate phenology.*

Response:

Thank you very much for your thorough review and valuable comments. We acknowledge that, at the current stage of this study, the core improvement of the WOFOST-EW model is indeed focused on enhancing the accuracy of phenology simulation by introducing the extreme weather function F(EW). We would like to emphasize that even though the current enhancement is limited to the phenological module, this strategy plays a critical role in improving the accuracy of crop yield simulation under extreme climate conditions. The rationale is as follows:

(1) Phenology is a key determinant of crop yield formation, and it serves as a major pathway through which extreme weather affects crop production. During the growth period from sowing to maturity, crops undergo several critical stages such as heading, flowering, and grain filling. The timing of these stages directly affects grain number and quality. Extreme weather events occurring within these sensitive phenological windows can cause irreversible damage. For example, extreme heat or drought during the grain filling stage may shorten the period, leading to insufficient filling and reduced grain weight, ultimately lowering the final yield (Chachar et al., 2016; Chowdhury et al., 2021; Ihsan et al., 2016; McMaster and Wilhelm, 2003). Frost events during heading or flowering can damage reproductive organs, significantly reducing the number of grains (Xiao et al., 2021). Similarly, prolonged waterlogging may delay phenological development or cause growth disruptions, making crops miss optimal development windows (Dickin and Wright, 2008; Wu et al., 2015).

(2) Existing crop models commonly face challenges in simulating phenology under extreme conditions. Although process-based crop models can simulate biomass accumulation and stress responses, their phenology modules are usually based on thermal time and photoperiod. Under extreme weather conditions, such formulations often fail to capture abnormal accelerations, delays, or interruptions in phenological development (Nóia Júnior et al., 2023). If phenology simulation is inaccurate, even a theoretically sound model of physiological processes will execute those processes at incorrect timings, introducing cumulative errors in biomass accumulation and yield prediction. Therefore, improving phenology prediction under extreme conditions is a critical step toward addressing the fundamental challenge of inaccurate crop modeling under climate extremes (Pei et al., 2025).

In summary, although the current version of the WOFOST-EW model does not extend modifications

to other physiological modules, the targeted enhancement of phenology alone has already significantly improved the model's robustness and yield prediction performance in extreme years. It also lays a solid foundation for future expansions to incorporate broader extreme weather response mechanisms. By integrating F(EW) via an LSTM framework, the model can learn and represent the nonlinear and dynamic influence of extreme weather on phenological development. This allows for a flexible and accurate correction of WOFOST's phenology predictions under extreme conditions, thus enabling the model to realistically capture the "developmental trajectory shift" experienced by crops under stress. We have elaborated on the rationale for focusing on phenology improvement in the revised manuscript's discussion section (see lines 320–368).

> *This approach attributes all extreme weather effects to phenology alone, which risks conflating physiological processes and may compromise the model's interpretability.*

Response:

Thank you for your thoughtful and constructive comments. We fully understand the reviewer's concern that attributing all extreme climate effects to phenological processes may potentially lead to confusion regarding physiological mechanisms and reduced model interpretability. Below, we provide a detailed response to these concerns:

(1) Regarding the concern about "conflating physiological processes":

We would like to clarify that we do not simply attribute all the effects of extreme climate to phenology. Rather, we explicitly identify and model one of the most direct, prominent, and yield-critical pathways—the phenological response of crops to extreme weather events. Numerous studies have shown that extreme heat, drought, or waterlogging events often first manifest in phenological disturbances, such as delayed heading, shortened grain filling duration, or premature maturity (Asseng et al., 2015; Chachar et al., 2016; Chen et al., 2014; Chowdhury et al., 2021; Ihsan et al., 2016; McMaster and Wilhelm, 2003; Sadras and Monzon, 2006; Tao and Zhang, 2013; Zahra et al., 2021). These disruptions significantly alter the sequence and duration of organ development, which in turn affects the biomass accumulation window and yield potential.

From a physiological perspective, extreme weather indeed directly affects fundamental processes such as photosynthesis, respiration, and water use. However, phenological changes are often a systemic reflection of the combined effects of these stresses. For example, high temperatures accelerate heat accumulation and enzyme activity related to photosynthesis, which promotes leaf area and chlorophyll growth and ultimately leads to earlier crop maturity and a shortened grain-filling period (Chen et al., 2014). When extreme heat induces premature senescence, or severe drought shortens the grain-filling phase, these are not isolated physiological responses but rather integrated expressions of the crop's response to environmental stress. These physiological responses are reflected at the phenological scale as a "shift in the developmental trajectory." Therefore, by

improving phenology simulation, the WOFOST-EW model indirectly captures the aggregated physiological stress responses at a macro scale. This is an integrative modeling approach that enhances model responsiveness under extreme conditions. Although WOFOST-EW does not explicitly modify parameters related to photosynthesis, it ensures that physiological processes such as biomass accumulation and yield formation are simulated within the correct phenological windows, which inherently account for the impact of extreme stress. In this way, physiological mechanisms are respected and indirectly represented through improved phenological timing.

To avoid confusion among physiological processes, we adopted a modular and isolated design strategy: extreme weather indicators are used only to drive the phenological module, while all other physiological processes follow the original WOFOST mechanisms and are not directly affected by the extreme weather variables. This preserves the structural independence of different physiological components and maintains clarity when interpreting the pathways through which various stresses operate. In fact, by improving phenological prediction, the model's "temporal alignment capability" is enhanced, which in turn improves the accuracy of simulating other physiological processes within the appropriate developmental stages.

(2) Regarding the concern about "compromise the model's interpretability":

We acknowledge that LSTM, as a deep learning method, lacks the mechanistic transparency of traditional physiological models. However, in terms of the model's overall behavioral interpretability, WOFOST-EW has not been weakened—on the contrary, its interpretive capacity has been enhanced in several key aspects:

- Enhanced interpretability under extreme weather conditions:
  The original WOFOST model may struggle with yield prediction under extreme conditions due to inaccuracies in phenology simulation. By adjusting phenological stages (e.g., simulating a shorter grain-filling period in hot years), WOFOST-EW aligns more closely with observed outcomes and provides better explanatory power regarding the causes of yield fluctuations.
- Improved user trust and applicability:
  Accurate simulation of phenological changes under extreme conditions increases the credibility of subsequent physiological processes, making the model more applicable in decision support and risk assessment contexts.

Since phenological accuracy is the foundation for reliable simulation of downstream processes such as biomass accumulation and yield formation, WOFOST-EW's improvements at the phenology level have substantially enhanced the model's predictive capability under extreme climate scenarios. This establishes a robust basis for more precise assessment of climate impacts and the development of adaptation strategies. Of course, we fully agree with the reviewer that future work should explore extending the influence of extreme weather to other physiological components of the WOFOST

model—such as direct effects on photosynthesis rate or water use efficiency—to achieve a more comprehensive representation of extreme stress mechanisms. This is a key direction for future development of the WOFOST-EW model.

We have added a detailed discussion of these points in the revised manuscript (see lines 320–368 and 422–438). Once again, we sincerely thank the reviewer for these constructive comments, which prompted us to further clarify the model's logic and scope of applicability.

LSTM training

● *The description of the LSTM network implementation lacks detail regarding the input sequence. Given that LSTM networks are designed for time series data and the extreme weather indices appear to be annual scalar values rather than time series over a growing season, it is unclear what sequence is fed into the LSTM. Is it a sequence of these annual scalars across the training period? This approach for LSTMs seems questionable, particularly when a held out test year falls within the training period and the output of the final timestep is taken as the prediction.*

Response:

We sincerely thank you for your critical and insightful comments regarding the implementation of the LSTM model. Your feedback is highly professional and extremely valuable to us. In response, we have provided a more detailed explanation of the LSTM modeling process in the Methods section (lines 164–178 of the revised manuscript), and would like to offer the following clarifications:

First, we clarify that the extreme weather indices fed into the LSTM network are not annual scalar values, but sequential time-series data calculated over the wheat growing season. The LSTM network processes these temporally dynamic sequences to capture the effects of extreme weather events on crop development, and uses them to estimate the time-varying function F(EW).

Furthermore, we acknowledge that our previous description of the LSTM methodology lacked clarity and may have led to confusion. Your concerns are completely valid and point to a critical issue of methodological rigor. Our specific training and testing strategies are as follows:

- WOFOST model calibration and validation:
  The WOFOST model was calibrated using historical data from 1980–2000, and subsequently validated using data from 2001–2020 to assess model performance across years.
- WOFOST-EW model training and testing:
  - Training and hyperparameter tuning (1980–2000):
    We used data from 1980 to 2000 as the training and internal validation set for the LSTM model. During this stage, we applied a leave-one-year-out cross-validation

(LOOCV) strategy and combined it with GridSearchCV for hyperparameter optimization and model training.

- Independent testing (2001–2020):
  After the LSTM model was trained and tuned using 1980–2000 data, we evaluated the final performance of the WOFOST-EW model using an entirely independent test dataset from 2001–2020.

In summary, all comparisons between the WOFOST and WOFOST-EW models in this study are based on performance over the independent test period (2001–2020). This ensures both fairness and the assessment of true predictive ability. We believe that these detailed clarifications regarding the input structure and validation strategy of the LSTM model will contribute to a better understanding of the methodological rigor of our manuscript.

● *It is also unclear what the training target of the LSTM is—does it aim to predict final yield, phenological stages, or something else? Furthermore, are the parameters of WOFOST kept fixed during the LSTM optimization? If the LSTM (and thus WOFOST-EW) is trained using phenological stage data while the original WOFOST model is not, then the comparison between the two models may not be fair or meaningful.*

Response:

We sincerely thank you for raising this critical question. Regarding the training objective of the LSTM model, parameter settings in WOFOST, and the fairness of model comparisons, we have provided further clarifications and revisions in the updated manuscript (see lines 164–178, 214–225, and Figure S1). Below is a detailed clarification:

First, the training objective of the LSTM model is not to directly predict yield or specific phenological dates. Rather, it is designed to estimate a data-driven extreme weather correction factor, F(EW), to better capture the nonlinear regulatory effects of extreme climate events on crop phenology. It is important to note that, in the baseline WOFOST, we used observational data on yield and phenology from 1980–2000 to calibrate the model. The parameter set that minimized the RMSE between simulated and observed yield and phenology was considered optimal. In WOFOST-EW, the same set of observational data was used to determine the extreme weather correction factor. Therefore, the comparison between the two models is fair.

Specifically, the LSTM takes as input multi-year sequences of extreme climate indices during the winter wheat growing season, and outputs the extreme weather correction factor. The LSTM does not replace the physiological modules of WOFOST; instead, it serves as an auxiliary model to correct the biases in phenology and yield predictions of the traditional WOFOST under extreme weather conditions. Thus, the optimization objective of LSTM is to ensure that when its output F (EW) variable is applied to the WOFOST model, the entire WOFOST-EW coupled model can

minimize the error between observed phenology and yield. We have made this clearer in lines 164–178 and 214–225 of the revised methods section.

Second, the process parameters in WOFOST were kept fixed throughout the entire experiment. That is, WOFOST-EW does not modify any physiological or process parameters of the WOFOST model during LSTM training. We ensured that WOFOST and WOFOST-EW shared identical inputs and configurations across all modules, including meteorological data, soil conditions, management practices, and crop-type parameters. The only difference is that WOFOST-EW applies the LSTM-derived F(EW) to correct the phenology module. This is explained in lines 214–225 of the revised manuscript.

Third, regarding your concern about the fairness of comparing the original WOFOST with WOFOST-EW, we believe our comparison approach is both fair and meaningful, for the following reasons:

- Calibration of the original WOFOST model: During the calibration stage (1980–2000), the original WOFOST model was optimized using both observed phenology (heading and maturity dates) and observed yield. This has been clearly stated in the methodology section (see lines 214–225). Therefore, the baseline WOFOST model represents a well-calibrated reference under typical climatic conditions.
- Enhancement in WOFOST-EW: The WOFOST-EW model is built upon this already-calibrated WOFOST baseline. The LSTM does not re-learn basic crop growth mechanisms from scratch; instead, it learns how to adjust the baseline model's biases in phenology and yield simulation under extreme weather conditions. The LSTM acts as a correction mechanism, not a standalone predictor.
- Evaluation on an independent test set: Most importantly, both models—WOFOST and WOFOST-EW—were evaluated on the same, fully independent test dataset (2001–2020). This ensures that neither model had prior exposure to the evaluation years during training or calibration. If WOFOST-EW shows significantly better performance on this test set—especially in extreme years—this serves as robust evidence that the incorporation of F(EW) via LSTM meaningfully improves predictive accuracy, lending strong scientific credibility and fairness to the comparison.

We have addressed these key points more thoroughly in Sections 2.3.2 to 2.3.4 of the methods, clarifying the LSTM training goal, the fixed nature of WOFOST parameters, and the fairness of the model comparisons. We hope this addresses your concerns and strengthens your confidence in the scientific rigor of our approach. Once again, we sincerely thank you for your thoughtful and constructive comments.

- *While WOFOST is calibrated using data from the years 1990–2000, the LSTM (and therefore*

WOFOST-EW) appears to be trained on a broader range of years through cross-validation. This introduces an unfair advantage in the evaluation.

Response:

We sincerely thank you for raising this critical question. Regarding the training objective of the LSTM model, the parameterization of WOFOST, and the fairness of comparisons between models, we have provided additional explanations in the revised manuscript (lines 164–178 and 214–225, as well as Fig. S1), and offer the following detailed clarifications:

First, the training objective of the LSTM model is not to directly predict crop yield or specific phenological dates, but to construct a data-driven extreme weather correction factor F(EW) that reflects the nonlinear regulatory effects of extreme climate events on crop phenology. Importantly, in the baseline WOFOST model, we used observational data from 1980–2000 on yield and phenology to calibrate the model. The optimal parameter set was determined when the root mean square errors (RMSE) between observed and simulated yield and phenological dates were minimized. In WOFOST-EW, the same observational data were used to estimate the correction factor F(EW). Therefore, the comparison between the two models is fair.

Specifically, the input to the LSTM consists of multiple years of extreme climate indices during the winter wheat growing season, and the output is the extreme weather correction factor. LSTM does not replace any physiological modules of WOFOST; instead, it serves as an auxiliary correction module to adjust the biases of the original WOFOST model under extreme weather conditions. Thus, the optimization objective of the LSTM is to ensure that, when the estimated F(EW) is applied to the WOFOST model, the coupled WOFOST-EW system minimizes the discrepancies between simulated and observed phenology and yield. We have provided more detailed improvements in the Methods section (lines 164–178).

Second, the physiological parameters of WOFOST remain unchanged throughout the entire experiment. During the LSTM training process, no crop physiological or process parameters were modified. We ensured that WOFOST and WOFOST-EW used identical input settings for all modules, including meteorological data, soil conditions, management practices, and crop parameter files. The only difference is that WOFOST-EW employs the LSTM-derived F(EW) factor to adjust the phenology module under extreme weather conditions.

Third, regarding your important concern about the fairness of comparing the original WOFOST and WOFOST-EW models, we emphasize that our comparison method is both fair and scientifically meaningful, for the following reasons:

- Calibration of the original WOFOST model:
  The baseline WOFOST model was calibrated using 1980–2000 data, including observed phenological dates (e.g., heading and maturity) and yield. This has been clearly stated in

the Methods section (see lines 214–225). Therefore, the WOFOST model was already optimized using actual observations to ensure its best possible performance under standard conditions.

- Enhancement in WOFOST-EW:
  The WOFOST-EW model builds upon this already calibrated baseline WOFOST. The LSTM model does not re-learn basic crop growth patterns from scratch, but instead learns how to adjust the original WOFOST predictions in the presence of extreme weather conditions. It serves as a corrective mechanism, not a standalone predictor.
- Independent testing dataset:
  Most importantly, both models were evaluated using the same independent test dataset (2001–2020), which was completely unseen during the training and calibration phases. This ensures that both models were assessed under the same conditions. If WOFOST-EW demonstrates significant improvements, particularly in extreme years, it provides strong evidence that the LSTM-derived F(EW) function offers a real enhancement in predictive capability, thus making the comparison both fair and scientifically valid.

We have elaborated on the LSTM training objective, the fixed WOFOST parameters, and the fairness of model comparisons in Sections 2.3.2–2.3.4 of the revised manuscript to address your concerns and reinforce the methodological rigor of the study. Once again, we sincerely appreciate your valuable suggestions.

● *It would be interesting to see a plot of only the intermediate LSTM output in the different years and if any patterns can be noticed w.r.t. the presence of extreme weather events.*

Response:

We sincerely appreciate your insightful suggestion. We have gladly adopted your valuable advice and have added a new figure in the revised manuscript to illustrate the interannual variation of the extreme weather function F(EW) learned by the LSTM model. This figure is intended to visually demonstrate how F(EW) responds to and quantifies the impact of extreme climate events.

The relevant content has been described in detail in lines 274–282 of the revised manuscript and is clearly presented in Figure 8.

Specifically, we plotted the annual distributions of extreme weather indices and the F(EW) values predicted by the LSTM model from 1980 to 2020. To more intuitively represent the magnitude of the negative impacts of extreme weather on crop growth, we used the metric 1 – F(EW): higher values of this indicator suggest a more pronounced effect of extreme weather on crop development in a given year.

Reproducibility

● *The provided codebase does not include any means to reproduce the results. For instance, the "LSTM" repository appears to be merely a clone of the Keras library.*

Response:

Thank you for your valuable comment regarding the issue of reproducibility. In response to your feedback, we have uploaded the core code used in this study during the revision process, including key function modules related to model construction, training, and prediction. We have also updated the manuscript to include information on how to access the code and details of its implementation. Please refer to lines 456–458 of the revised manuscript for further information.

● *Section 2.2.2 states that yield data is used from the Agricultural Yearbook of the respective provinces, and is supplemented with data not disclosed by survey bureaus. Where does this supplemented data come from*

Response:

Thank you for your concern regarding the accuracy of data sources. We would like to clarify the issue you raised about the non-public data: all crop yield data used for model training and validation in this study were obtained from officially published statistics in the Agricultural Yearbooks of the respective provinces. We acknowledge that our previous wording was not sufficiently rigorous. We have corrected and clarified the relevant statements in the revised manuscript. Please refer to lines 122–124 for details.

Calibration WOFOST

● *It is unclear which parameters are selected during the WOFOST calibration process. How many? How many iterations? Which (if any) parameter value ranges were used?*

Response:

We sincerely thank the reviewer for your attention to the parameter calibration process of the WOFOST model. We have revised and improved this section in the manuscript to enhance the transparency and reproducibility of the methods. Specifically, we first conducted a sensitivity analysis to identify the key parameters, during which both first-order and total sensitivity indices were calculated. Based on these results, the 10 parameters with the highest global sensitivity were selected for calibration.

For parameter optimization, we adopted the widely used Shuffled Complex Evolution-University of Arizona (SCE-UA) algorithm, which has been extensively applied in crop model inversion. We set 200 iterations to ensure the stability and convergence of the results. The value ranges of each parameter were determined based on relevant literature and the physiological characteristics of the crop. The detailed parameter ranges are listed in Tables S3 and S4 of the supplementary material.

The above details have been added in lines 214–225 of the revised manuscript and further elaborated in Texts S1 and S2 of the supplementary material. Once again, we thank you for your valuable comments, which have helped improve the completeness and clarity of our methodological description.

Agricultural management data

● *No details are provided regarding farm management. The authors should clarify whether the winter wheat was irrigated and specify the fertilization applied. Phenological stages*

Response:

Thank you very much for your valuable suggestion. We have added detailed information on field management practices in the revised manuscript, including irrigation and fertilization for winter wheat. Please refer to lines 110–121 of the revised manuscript for the updated content.

● *The authors evaluate two phenological stages, namely heading and maturity. In WOFOST, phenology is represented by the Development Stage (DVS), which ranges from 0 (emergence) to 2 (maturity), with 1 corresponding to flowering. What DVS value corresponds to the heading stage?*

Response:

Thank you for raising this important question. In the original WOFOST model, the phenological development stage (DVS) is defined with 0 representing emergence, 1 for flowering, and 2 for maturity. However, in this study, the observed phenological data used for evaluation primarily include heading and maturity stages. Given that heading and flowering occur nearly simultaneously in the major winter wheat production regions of China, we made an appropriate adjustment to the representation of phenological stages in our model.

Specifically, to better reflect the phenological characteristics of winter wheat in our study area and meet simulation needs, we redefined the thermal time T_i as the sum of effective temperatures between emergence and heading, or between heading and maturity. Accordingly, the DVI (development stage index) was reset such that DVI = 1 corresponds to heading, and DVI = 2 to maturity.

This clarification has been added to the revised manuscript at lines 194–197.

Other

● *Abstract: There is no mention of why deep learning models should be expected to help improve winter wheat simulations in this setting.*

Response:

Thank you for your valuable suggestion. We have revised the abstract to explicitly state the

motivation and advantages of incorporating a deep learning model. The relevant content has been updated in the revised manuscript at lines 18–20 of the abstract.

● *Section 2.3.2: LDD and HDD terms are used without introduction*

Response:

Thank you for your valuable suggestion. We have added the definitions of the indices in the revised manuscript at lines 135–137.

● *Section 2.3.3: LSTM is a recurrent neural network architecture that mitigates the inherent weakness in rnns in dealing with long temporal dependencies. It is not inherently stable and high performing in long term prediction tasks.*

Response:

Thank you for your valuable suggestion. We have revised the definition of LSTM accordingly. Please refer to lines 164–168 in the revised manuscript.

● *Section: 2.3.4: What are the parameter values used for t_base and t_max in the WOFOST temperature response functions? (F(T) in the paper). Discussion on the behavior and possible limitations could be beneficial.*

Response:

Thank you for your valuable suggestion—this issue is indeed critical to the model's accuracy and adaptability. We have addressed it by explicitly stating the parameter settings for the temperature response function F(T) in lines 198–200 of the revised manuscript. Specifically, the base temperature (T_b) and the upper temperature threshold (T_m) are set to 0 °C and 30 °C, respectively.

Additionally, we have included a new discussion on the limitations of the F(T) function in lines 304–319. In particular:

- F(T)'s inability to capture the suppressive effects of extreme heat: Under the current formulation, when the temperature exceeds T_m, the value of F(T) remains constant. This means the model no longer responds to further increases in temperature, and therefore fails to account for physiological stress responses such as reduced photosynthetic efficiency, accelerated respiration, or reproductive organ damage. As a result, the model may systematically overestimate growth potential under extreme heat conditions.
- Simplified representation of cold stress: F(T) is set to zero when the temperature drops below T_b. While this reflects the concept of "growth cessation," it does not differentiate between varying levels of cold stress, such as mild chilling versus severe frost damage, which may lead to underrepresentation of the biological consequences of low temperatures.

To address these limitations, we introduced the extreme weather response function F(EW) in this

study, as a complement to the traditional F(T). F(EW) enables a more comprehensive representation of phenological disturbances caused by extreme weather conditions. Once again, thank you for your thoughtful suggestion—it has contributed to improving the conceptual soundness and direction of model development.

● *Tables S3 and S4: Would it be possible to include error bars/confidence intervals?*

Response:

Thank you very much for your suggestion. The 95% confidence intervals have now been added to both tables. Please refer to the revised Table S5 and Table S6 in the revised manuscript.

● *The separate evaluation in years with extreme weather events is very interesting. Maybe including a statistical significance test could help in giving insight to whether two years of data is sufficient to claim an improvement.*

Response:

Thank you very much for your valuable suggestion. To further enhance the rigor of our findings, we have incorporated your recommendation and added statistical significance tests. For detailed results, please refer to lines 251–257, 274–282, and 296–301 in the revised manuscript.

References

Al Issawi, M., Rihan, H. Z., El Sarkassy, N., and Fuller, M. P.: Frost Hardiness Expression and Characterisation in Wheat at Ear Emergence, J. Agron. Crop Sci., 199, 66-74, 2013.

Asseng, S., Ewert, F., Martre, P., Rötter, R. P., Lobell, D. B., Cammarano, D., Kimball, B. A., Ottman, M. J., Wall, G. W., and White, J. W.: Rising Temperatures Reduce Global Wheat Production, Nat. Clim. Chang., 5, 143-147, 2015.

Chachar, M. H., Chachar, N. A., Chachar, Q., Mujtaba, S. M., Chachar, S., and Chachar, Z.: Physiological Characterization of Six Wheat Genotypes for Drought Tolerance, International Journal of Research–Granthaalayah, 4, 184-196, 2016.

Chen, J., Tian, Y., Zhang, X., Zheng, C., Song, Z., Deng, A., and Zhang, W.: Nighttime Warming Will Increase Winter Wheat Yield through Improving Plant Development and Grain Growth in North China, J. Plant Growth Regul., 33, 397-407, 2014.

Chowdhury, M. K., Hasan, M. A., Bahadur, M. M., Islam, M. R., Hakim, M. A., Iqbal, M. A., Javed, T., Raza, A., Shabbir, R., Sorour, S., Elsanafawy, N. E. M., Anwar, S., Alamri, S., Sabagh, A. E., and Islam, M. S.: Evaluation of Drought Tolerance of some Wheat (Triticum Aestivum L.) Genotypes through Phenology, Growth, and Physiological Indices, in: Agronomy, edited10.3390/agronomy11091792, 2021.

Colmer, T. D. and Greenway, H.: Ion Transport in Seminal and Adventitious Roots of Cereals During O2 Deficiency, J. Exp. Bot., 62, 39-57, 2011.

Colmer, T. D. and Voesenek, L.: Flooding Tolerance: Suites of Plant Traits in Variable Environments, Funct. Plant Biol., 36, 665-681, 2009.

Dickin, E. and Wright, D.: The Effects of Winter Waterlogging and Summer Drought on the Growth and Yield of Winter Wheat (Triticum Aestivum L.), Eur. J. Agron., 28, 234-244, https://doi.org/https://doi.org/10.1016/j.eja.2007.07.010, 2008.

Ebrahimi-Mollabashi, E., Huth, N. I., Holzwoth, D. P., Ordóñez, R. A., Hatfield, J. L., Huber, I., Castellano, M. J., and Archontoulis, S. V.: Enhancing APSIM to Simulate Excessive Moisture Effects on Root Growth, Field Crops Res., 236, 58-67, 2019.

Fuller, M. P., Fuller, A. M., Kaniouras, S., Christophers, J., and Fredericks, T.: The Freezing Characteristics of Wheat at Ear Emergence, Eur. J. Agron., 26, 435-441, 2007.

Harrison, M. T.: Climate Change Benefits Negated by Extreme Heat, Nat. Food, 2, 855-856, https://doi.org/10.1038/s43016-021-00387-6, 2021.

Herzog, M., Striker, G. G., Colmer, T. D., and Pedersen, O.: Mechanisms of Waterlogging Tolerance in Wheat–a Review of Root and Shoot Physiology, Plant Cell Environ., 39, 1068-1086, 2016.

Ihsan, M. Z., El-Nakhlawy, F. S., Ismail, S. M., Fahad, S., and Daur, I.: Wheat Phenological Development and Growth Studies as Affected by Drought and Late Season High Temperature Stress Under Arid Environment, Front. Plant Sci., 7, 795, 2016.

Jitsuyama, Y.: Hypoxia-Responsive Root Hydraulic Conductivity Influences Soybean Cultivar-Specific Waterlogging Tolerance, American Journal of Plant Sciences, 8, 770, 2017.

Kaur, G., Singh, G., Motavalli, P. P., Nelson, K. A., Orlowski, J. M., and Golden, B. R.: Impacts and Management Strategies for Crop Production in Waterlogged or Flooded Soils: A Review, Agron. J., 112, 1475-1501, 2020.

Kotula, L., Clode, P. L., Striker, G. G., Pedersen, O., Läuchli, A., Shabala, S., and Colmer, T. D.: Oxygen Deficiency and Salinity Affect Cell-Specific Ion Concentrations in Adventitious Roots of Barley (H Ordeum Vulgare), New Phytol., 208, 1114-1125, 2015.

Li, X., Cai, J., Liu, F., Dai, T., Cao, W., and Jiang, D.: Cold Priming Drives the Sub-Cellular Antioxidant Systems to Protect Photosynthetic Electron Transport Against Subsequent Low Temperature Stress in Winter Wheat, Plant Physiol. Biochem., 82, 34-43, 2014.

Li, Y., Hou, R., and Tao, F.: Interactive Effects of Different Warming Levels and Tillage Managements on Winter Wheat Growth, Physiological Processes, Grain Yield and Quality in the North China Plain, Agric. Ecosyst. Environ., 295, 106923, https://doi.org/https://doi.org/10.1016/j.agee.2020.106923, 2020.

Liu, K., Harrison, M. T., 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 Future, 8, e2020EF001801, 2020.

Liu, L., Xu, H., Liu, S., and Liu, X.: China'S Response to Extreme Weather Events Must be Long Term, Nat. Food, 4, 1022-1023, https://doi.org/10.1038/s43016-023-00892-w, 2023.

Mcmaster, G. S. and Smika, D. E.: Estimation and Evaluation of Winter Wheat Phenology in the Central Great Plains, Agric. For. Meteorol., 43, 1-18, https://doi.org/https://doi.org/10.1016/0168-1923(88)90002-0, 1988.

Mcmaster, G. S. and Wilhelm, W. W.: Phenological Responses of Wheat and Barley to Water and Temperature: Improving Simulation Models, The Journal of Agricultural Science, 141, 129-147, 2003.

Nguyen, T., Son, S., Jordan, M. C., Levin, D. B., and Ayele, B. T.: Lignin Biosynthesis in Wheat (Triticum Aestivum L.): Its Response to Waterlogging and Association with Hormonal Levels, BMC Plant Biol., 16, 1-16, 2016.

Nóia Júnior, R. D. S., Asseng, S., García-Vila, M., Liu, K., Stocca, V., Dos Santos Vianna, M., Weber, T. K. D., Zhao, J., Palosuo, T., and Harrison, M. T.: A Call to Action for Global Research on the Implications of Waterlogging for Wheat Growth and Yield, Agric. Water Manag., 284, 108334, https://doi.org/https://doi.org/10.1016/j.agwat.2023.108334, 2023.

Nyaupane, S., Poudel, M. R., Panthi, B., Dhakal, A., Paudel, H., and Bhandari, R.: Drought Stress Effect, Tolerance, and Management in Wheat–a Review, Cogent Food Agr., 10, 2296094, 2024.

Pei, J., Tan, S., Zou, Y., Liao, C., He, Y., Wang, J., Huang, H., Wang, T., Tian, H., Fang, H., Wang, L., and Huang, J.: The Role of Phenology in Crop Yield Prediction: Comparison of Ground-Based Phenology and Remotely Sensed Phenology, Agric. For. Meteorol., 361, 110340, https://doi.org/10.1016/j.agrformet.2024.110340, 2025.

Porter, J. R. and Gawith, M.: Temperatures and the Growth and Development of Wheat: A Review, Eur. J. Agron., 10, 23-36, 1999.

Sadras, V. O. and Monzon, J. P.: Modelled Wheat Phenology Captures Rising Temperature Trends: Shortened Time to Flowering and Maturity in Australia and Argentina, Field Crops Res., 99, 136-146, 2006.

Shroyer, J. P., Mikesell, M. E., and Paulsen, G. M., (Eds.): Spring Freeze Injury to Kansas Wheat, Cooperative Extension Service, Kansas State University, 1995.

Tao, F. and Zhang, Z.: Climate Change, Wheat Productivity and Water Use in the North China Plain: A New Super-Ensemble-Based Probabilistic Projection, Agric. For. Meteorol., 170, 146-165, 2013.

Tao, F., Rötter, R. P., Palosuo, T., Díaz-Ambrona, C. G. H., Mínguez, M. I., Semenov, M. A., Kersebaum, K. C., Nendel, C., Cammarano, D., and Hoffmann, H.: Designing Future Barley Ideotypes Using a Crop Model Ensemble, Eur. J. Agron., 82, 144-162, 2017a.

Tao, F., Xiao, D., Zhang, S., Zhang, Z., and Rötter, R. P.: Wheat Yield Benefited from Increases in Minimum Temperature in the Huang-Huai-Hai Plain of China in the Past Three Decades, Agric. For.

Meteorol., 239, 1-14, 2017b.

Wu, X., Tang, Y., Li, C., Wu, C., and Huang, G.: Chlorophyll Fluorescence and Yield Responses of Winter Wheat to Waterlogging at Different Growth Stages, Plant Prod. Sci., 18, 284-294, 2015.

Xiao, L., Liu, B., Zhang, H., Gu, J., Fu, T., Asseng, S., Liu, L., Tang, L., Cao, W., and Zhu, Y.: Modeling the Response of Winter Wheat Phenology to Low Temperature Stress at Elongation and Booting Stages, Agric. For. Meteorol., 303, 108376, https://doi.org/https://doi.org/10.1016/j.agrformet.2021.108376, 2021.

Zahra, N., Wahid, A., Hafeez, M. B., Ullah, A., Siddique, K. H. M., and Farooq, M.: Grain Development in Wheat Under Combined Heat and Drought Stress: Plant Responses and Management, Environ. Exp. Bot., 188, 104517, https://doi.org/https://doi.org/10.1016/j.envexpbot.2021.104517, 2021.