Response to Reviewers

Preface

During the discussion stage, the regional resolution of the crop emulator has been updated to conform with the upcoming OSCAR v4, and we have updated some results of this manuscript accordingly. In addition, some metrics used to compare and examine the performance of the crop emulator are changed. For example, in Fig. 4, instead of using RMSE, we now use RRMSE to evaluate the differences between *firr* and *noirr* crops. In Section 4.1, we replaced Pearson correlation coefficients with RRMSE to focus on the overall alignment between the original and emulated results, rather than only on the trend. Despite the changes, the overall workflow and main conclusions of this paper remain robust.

We thank the reviewer for the thoughtful comments, which have helped us improve the manuscript. In response, we have made several revisions. All changes are documented in our point-by-point response below. We have carefully addressed all reviewer comments, with our responses highlighted by underlines as detailed below.

Reply to RC1

In this manuscript, the authors present a crop yield emulator of the Agricultural Model Intercomparison and Improvement Project (AgMIP) global gridded crop models (GGCMs). Their crop yield emulator was designed to be driven by CTWN output from OSCAR for novel scenarios, but was trained on and validated using publicly available model intercomparison data from GGCMs. The authors describe the model development, validation with out-of-sample GGCM results, as well as manipulative field experiments. There is a well-established need for crop yield emulators, and it is exciting to see this field growing. However, as it currently stands, I think revisions are needed to improve the clarity and readability of the manuscript. Line by line questions are included below but one overarching comment is that it is not particular clear what and how the emulator is actually connected with/related to OSCAR.

Isn't OSCAR now on version > V3 (Gasser et al. 2020)? How does OSCAR-crop v1.0 relate to it? Is the crop emulator standalone from OSCAR, which is why it is only on v1? Will it be included in a future OSCAR release? Some clarity on whether the emulator is a module/component of OSCAR or fully independent, that is, soft-coupled with OSCAR, would be helpful. Is there a specific version of OSCAR that the emulator is compatible with? Or is it also backwards compatible with the V1 OSCAR release? These sorts of details, and the possibility of coupling the crop emulator with other RCMs, would be helpful to readers and potential users. Were any OSCAR driven emulation results included in manuscript?

Thank you for these important questions regarding the relationship between OSCAR-crop v1.0 and the OSCAR model. We clarify these points below:

Version Relationship & Naming:

- The most recent version of OSCAR is v3.3.
- OSCAR-crop v1.0 is the first released version of our new crop emulator module. Its code structure is fully compatible with OSCAR v3.3. The version number refers to the emulator itself, not OSCAR.

Coupling Status & Compatibility:

- Standalone Use: the crop emulator can be run independently, as was done for all analyses in this manuscript.
- Integration with OSCAR: it is designed as a potential module for OSCAR. While not yet merged into the main OSCAR branch, it can be seamlessly coupled with OSCAR v3.3 due to a shared code framework.
- Backward Compatibility: due to a major structural overhaul of OSCAR since version 3, the emulator is not compatible with OSCAR v2 or earlier.

Coupling with Other Models:

Coupling with other Reduced-Complexity Models (RCMs) or Integrated Assessment Models (IAMs) is conceptually straightforward. It requires that the host model can provide the necessary climate drivers (temperature, precipitation, CO_2) on a compatible regional scale. The emulator's standalone design facilitates this.

Results in This Manuscript:

The results presented in this manuscript are from offline, standalone emulator simulations. We did not use OSCAR-driven climate projections in this study. Our focus was on emulating and validating against the ISIMIP3 GGCM ensemble. The coupling with OSCAR is a readiness feature for future applications.

To provide the precise technical details requested (compatibility, coupling instructions, version specifics), we will ensure they are fully documented in the emulator's README file and code repository. We will add a brief note in the manuscript's 'Code Availability' section to direct users to the repository for complete documentation.

L26 : "to estimate yield responses under various scenarios" is this under various future climate scenarios? Or are socioeconomic conditions also part of the prediction process?

Yes, the responses are estimated under historical, SSP126, SSP370, and SSP585 climate scenarios. The socio-economic conditions including the nitrogen inputs are fixed for the GGCM simulations in ISIMIP3.

The corresponding text is changed to:

"These crop models used bias-corrected historical and future (SSP126, SSP370, and SSP585) climate scenarios under fixed human direct forcing to estimate yield responses to *C*, *T* and *W*."

L33: "bridging the gap between complex crop models and statistic models" what do the authors mean by this gap?

We position our model between complex process-based models, which are computationally expensive and data-heavy, and statistical models, which can be overly simplistic. Our goal is to provide a balanced alternative that captures essential crop dynamics more efficiently.

The corresponding text is changed into:

"providing a middle-ground between complex crop models and statistic models".

L43-48: In the chunk of text starting with "In contrast" and ending with "(Folberth et al., 2025)". As it currently reads, with where manuscripts are cited, it seems like the only other existing crop yield emulator is documented in Abramoff et al., 2023, but other crop yield emulators exist. The authors should cite more than one other crop yield emulator in this section or clarify why this emulator is so unique among them.

Thanks for the suggestions. More references are added in main text.

L56 - 60: In this section of text, the authors have been discussing a mix of crop emulators and crop yield simulations generated by the complex crop models. The second half of the paragraph is hard to follow because it is unclear what type of model is being discussed. For example, with the sentence "Despite the wide range of outcomes due to different model structures, parameterization schemes, calibration processes and input data quality (Folberth et al., 2019;Müller et al., 2024), these projections exhibit reduced uncertainty for rice and soybean and enhanced robustness for maize and wheat (Jägermeyr et al., 2021)" are the authors referring to the complex crop models participating in GGCMI Phase 3 as having a wide range of outcomes? Or were these national crop yield emulators that the authors are building up with this work by developing a sub-national crop emulator?

Thanks for highlighting the lack of clarity in this paragraph. We agree that the distinction between GGCMs and crop emulators should be made explicit. To improve readability, we will revise the text in lines 56-60 to clearly specify the subject.

The revised text is:

"Despite the wide range of outcomes due to different model structures, parameterization schemes, calibration processes and input data quality (Folberth et al., 2019; Müller et al., 2024), the projections in GGCMI Phase 3 exhibit reduced uncertainty for rice and soybean and enhanced robustness for maize and wheat (Jägermeyr et al., 2021)."

L74: "It emulates crop yields at a national level for most countries, with sub-national outputs in six large-area countries (Australia, Brazil, Canada, China, Russia, and the USA)." How many regions in total? Is this enough to be considered subnational modeling capabilities? As described in L60?

Following the discussion period, the emulator has been updated to align with the OSCAR v4 regional aggregation scheme, bringing the total to 311 regions. This includes sub-national modeling for six large-area countries (Australia, Brazil, Canada, China, Russia, and the USA) while maintaining national-level resolution for others. We believe this represents a robust sub-national modeling capability, as it captures internal spatial heterogeneity within the world's largest agricultural producers, which collectively account for a significant portion of global crop land and climatic variance.

The revised text is:

"The emulator encompasses 311 regions in total, providing sub-national modelling capabilities for six large-area countries—Australia, Brazil, Canada, China, Russia, and the USA—to capture internal spatial heterogeneity. For the remainder of the globe, the model operates at a national level."

L99: "The input variables provided in the repository", what do you mean by the input variables, are you referring to the ISIMIP data that the crop emulator will use as inputs? Or are these data included in the emulator repository for emulator users?

The term "input variables" in this context refers to the environmental and management drivers sourced from the ISIMIP repository to drive the complex crop models. Section 2.2 details the preprocessing of this raw ISIMIP data into a structured format suitable for training the emulator.

While the full raw ISIMIP dataset is hosted externally, we provide the processed training datasets directly within the emulator repository to facilitate immediate use and reproducibility.

We have added the term "raw" to the manuscript (e.g., "raw input variables") to more clearly distinguish between the original ISIMIP source data and the processed data.

Equations (3 & 4): Where do the weights between the regional climate and crop-specific/regional growing season crop come from? Is Oscar producing the growing-season regional temperatures and precipitation?

As described in Section 2.2, the weights used in Equations 3 and 4 are derived from crop calendar and land-use data. The former specifies the duration between planting and harvesting dates for each region and crop type, the latter indicates crop-specific cropland areas. These weights effectively aggregate climate variables into crop-specific/ regional growing-season averages.

Regarding the model's output: In this study, the crop emulator is run independently of OSCAR. We implemented the regional climate module (Equations 1 and 2) directly within the crop emulator using updated parameters. And the growing-season climate can only be generated by the crop emulator not OSCAR at this stage.

The following text is added before the two equations:

"The growing season variables ($\Delta T_{gs}^{i,c}$ and $\Delta P_{gs}^{i,c}$) are calculated following the precprocessing steps described in Section 2.2."

L145 - Why would the concatenation matter? If doing global to regional linear pattern scaling?

Thank you for this insightful question. You are correct that under the assumptions of global-to-regional pattern scaling—where the relationship is linear and time-invariant—the concatenation scheme should not, in principle, matter.

Our decision to test this was primarily a precautionary check. While pattern scaling assumes linearity, we wanted to empirically verify that our fitting results were not inadvertently influenced by how we handled transitions between different scenarios (historical and SSPs) in the concatenated time series. The sensitivity analysis in Figs. S2–S5 confirms your theoretical expectation: the choice of concatenation scheme has only minor effects on the fitted parameters, which supports both the validity of the linear assumption and our use of the simpler direct merging approach.

L175 - Is a consistent functional form used across crop types? Or is it the best emulator per region x crop type?

Thank you for your question. The functional forms differ by crop type and region for climate impacts, allowing the emulator to capture region- and crop-specific responses to temperature and precipitation. For nitrogen impacts, due to the limited data, we apply a consistent functional form across all crop types and regions.

Figure 3: Is the y-axis RCCO2? Or is it RC? It might be helpful to include that labeling on the y-axis

Thank you for pointing this out. The y-axis represents $R^{c}_{CO_2}$, and we have modified the figure to make this labeling clearer in the revised manuscript.

Equation 9, which subtracts the perception pi control from the climate scenario, appears inconsistent with the relative perception changes described above.

Thank you for pointing this out. Equation 9 defines ΔP^{c}_{gs} as the difference between scenario precipitation and the pi control reference. The input variable used in the crop emulator is the relative change $\Delta P^{c}_{gs}/\Delta P^{c}_{gs,\,pi}$, which is consistent with the description of relative precipitation changes in the text.

~ L340 For the N fertilizer effect, could the authors clarify whether they are conducting the field experiments following the van Grinsven et al. (2022) or if they are using data published from this field experiment? Given the data limitations and the assumptions that had to be made, is it necessary to include this term in the emulator?

No field experiment is conducted, and we use the published data from the van Grinsven et al. (2022). The reason for this term is to strengthen the emulator's ability to reproduce real-life crop yields. And the comparison with FAO data further proves the importance of including this term.

To make it clearer, we have revised the text as follows:

"To facilitate nitrogen impact assessment in the crop emulator, data from long-term field experiments are employed to quantify crop yield responses to nitrogen inputs following van Grinsven et al. (2022)".

In Figure 9: What do the symbol makers indicate? Is it the distribution of the global average of the sub-national absolute differences? Or is it the sub-national absolute differences?

The makers indicate the relative differences between the emulated global crop yields and the original ISIMIP global crop yields.

We have modified the title as follows:

"Yield differences between crop emulator and original ISIMIP3a historical simulations shown as: global maps of GGCM-averaged sub-national absolute differences (tDM ha^{-1}) and violin plots of relative differences (%) for global average values across crop-GGCM combinations. For each crop in the violin plots, the markers along the horizontal axis follow the alphabetic order of the eight GGCMs."

Section 5.1 is difficult to follow. Are the authors comparing emulator results with experimental observations? Is the emulator being used to predict the experimental change in yield response? Or are the field experiment results being incorporated into the emulator by "ground[ing] the emulator's projections in real-world experimental evidence"? Furthermore, it is not entirely clear how the MC relates to the observational/field experiments.

Thank you for these helpful comments. We appreciate the opportunity to clarify this section.

Purpose of the comparison:

Yes, we are comparing emulator results with experimental observations (from FACE, OTC, and field warming experiments). The goal is to evaluate whether our emulator can reproduce the distributions of crop yield sensitivities observed in real-world experiments, thereby grounding the emulator's projections in empirical evidence.

Relationship between emulator and experiments:

The field experiment results are not incorporated into the emulator. Instead, we use them as an independent benchmark for validation. The emulator's ability to reflect real-world situations stems from its foundation in the original Global Gridded Crop Models (GGCMs), which themselves are calibrated on observational data from field experiments and other sources. Thus, the consistency we observe between emulator outputs and experimental results is encouraging but not entirely surprising—it indicates that the emulator successfully preserves the core response patterns of the underlying process-based models.

Role of Monte Carlo:

The MC ensemble is used to generate a probabilistic distribution of yield sensitivities, reflecting parametric uncertainty in the emulator. By running 1000 parameter configurations, we capture a range of plausible yield responses. This allows us to compare not just the mean emulator response to experiments, but also the spread of uncertainty—which, as noted in Fig. 10, is narrower than the spread observed in field experiments, suggesting that the GGCMs may not fully capture the range of real-world variability.

We have modified the first sentence of Section 5 accordingly:

"To ground the emulator's projections in real-world experimental evidence, distributions of crop yield sensitivities derived from FACE, OTC, and field warming experiments are compared with diagnostic outputs from the crop emulator."

Reply to RC2

This study develops a model that designs, calibrates, and validates a crop yield emulator and integrates it as a module within the compact Earth system model or simple climate model OSCAR. The model design and intended use are clearly presented, featuring national and six sub national resolutions, annual temporal resolution, and two modes of operation, thus fitting well within the scope of Geoscientific Model Development. In addition, the calibration framework explicitly specifies a decomposition of response functions with four drivers CO2, growing season temperature, precipitation or water availability, and nitrogen, and a selection procedure based on statistical criteria, while validation is extended to comparisons against ISIMIP3b, ISIMIP3a, experimental evidence including FACE, OTC, and warming experiments, and FAO yield statistics. Nevertheless, because the authors should further strengthen the clarity of model structure and assumptions, the completeness of the reproducibility package, and the independence and generalizability of the validation, I provide the following major comments.

1. The model decomposes total yield response as the product of responses to CO2, temperature, precipitation, and nitrogen as in Eq. 6. While this structure offers advantages in computational efficiency and interpretability, it may fundamentally weaken interaction terms and nonlinear couplings. Did the authors assess how much interaction remains in the ISIMIP3b and ISIMIP3a data, for example via residual structure or systematic biases in particular regimes. If not, the manuscript would be more convincing if it included, even briefly, error distributions in regimes where interactions are expected to be large such as high CO2, high temperature, and dry conditions, or a simple comparison of the performance and complexity trade off when allowing selective interaction terms instead of a purely multiplicative decomposition.

   Thank you for this thoughtful and important question. You raise a valid point about the potential loss of interaction effects in multiplicative decomposition. We would like to clarify that interaction terms are indeed preserved in our formulation, though implicitly rather than explicitly.

   To illustrate, consider a simplified version of Eq. 6 with only two drivers: CO_2 and growing season temperature (T_{gs}):

$$R^c = R^c_{CO_2} \times R^c_{T_{gs}}$$

   The partial derivative with respect to CO_2 is:

$$\frac{\partial R^c}{\partial CO_2} = \frac{\partial R^c_{CO_2}}{\partial CO_2} \times R^c_{T_{gs}} + \frac{\partial R^c_{T_{gs}}}{\partial CO_2} \times R^c_{CO_2}$$

   Since $R^c_{T_{gs}}$ is not a function of CO_2, the second term vanishes, leaving:

$$\frac{\partial R^c}{\partial CO_2} = \frac{\partial R^c_{CO_2}}{\partial CO_2} \times R^c_{T_{gs}}$$

   Assuming a functional form is fitted, then:

$$\frac{\partial R^c_{CO_2}}{\partial CO_2} \neq 0$$

   The marginal effect of CO_2 on yield depends on $R^c_{T_{gs}}$, which varies with temperature. This dependence constitutes an interaction between CO_2 and temperature—the effect of one driver changes with the level of the other.

More generally, in a multiplicative model $R = f(C) \times g(T)$, the effect of C on R is $f'(C) \times g(T)$, which depends on T. Thus, interactions arise naturally from the multiplicative structure whenever the component functions are nonlinear.

That said, we acknowledge that the specific form of these interactions is constrained by the functional forms chosen for each driver.

To assess whether the emulator captures the range of interactions present in the original GGCMs, we compare end-of-century yield differences between our emulator and the ISIMIP3b ensemble under the high-emission SSP585 scenario, a regime where interactions are expected to be large. These comparisons are presented in Figures S11–S18. The generally good agreement suggests that our multiplicative structure adequately reproduces the interaction patterns embedded in the underlying process-based models.

We have added one paragraph to the end of Section 4.1:

"These results also illustrate the interaction effects inherent in our multiplicative model structure. The SSP585 scenario represents a high-stress regime where elevated CO_2, high temperatures, and potential drought conditions co-occur—precisely the conditions where interactions between drivers are expected to be largest. The good agreement between emulator and GGCM outputs under this extreme scenario indicates that the implicit interactions arising from our nonlinear component functions capture the coupled responses present in the original process-based models."

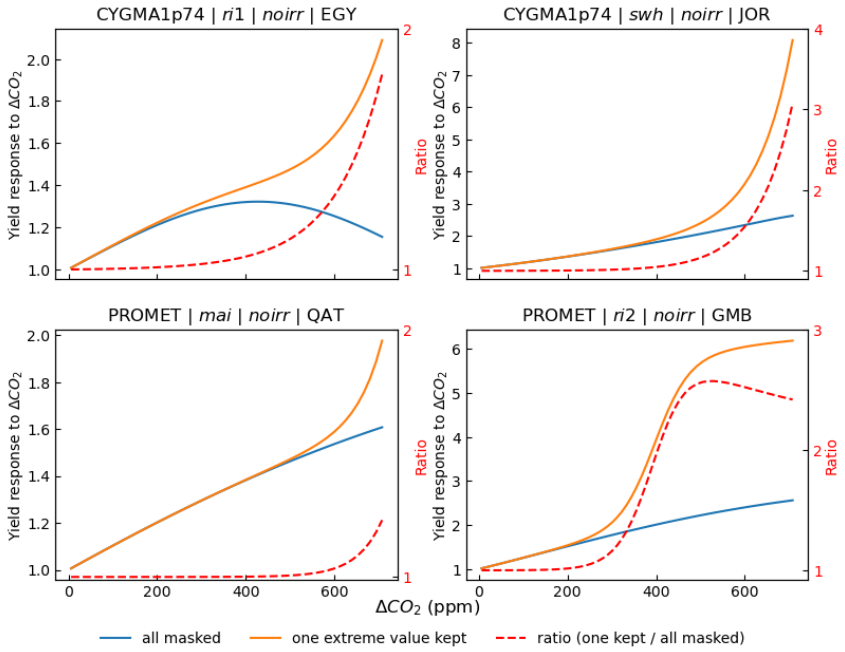
2. The authors select functional forms using R squared and BIC and explore a large number of combinations across regions, crops, irrigation, and drivers. Did the authors evaluate via some form of cross validation how stable the selected functional forms are across scenarios such as different SSPs and time periods near future versus end of century. Although the statement that BIC helps suppress overfitting is reasonable, it is important to more clearly demonstrate robustness using an explicit train and validation split.

Thank you for this important question. In our initial calibration, we implemented a train-validation split: the training dataset comprised historical, SSP126, and SSP585 scenarios, while the validation dataset used the SSP370 scenario. However, we encountered a limitation: SIMPLACE-LINTUL5—one of the GGCMs in our ensemble—does not report simulation results under SSP370, which would have led to inconsistent treatment across models.

We therefore compared the BIC-based selection approach against the train-validation split for the remaining GGCMs and found that both methods yielded similar functional form selections and performance metrics. Given this consistency, and to avoid discarding valuable information from the full dataset, we adopted the BIC approach for all calibrations. This allows us to retain more data during fitting while BIC's penalty term on model complexity provides inherent protection against overfitting.

3. The manuscript states that extreme value regions represent only 0.02 percent of the dataset but can distort aggregated results, so masking is applied. At the same time, the parameters for those regions are retained, reflecting the original model, and the choice is left to the user. The authors also list the number of extreme regions for specific GGCM, crop, and irrigation combinations. In that case, could the authors provide quantitative examples, a sensitivity analysis, showing which regions drive the differences and by how much, and how global and regional results change depending on whether such regions are included or excluded.

Thank you for this helpful suggestion. We have now added a sensitivity analysis in the supplementary Text S1 to quantify the influence of extreme-response regions on aggregated results.



Specifically, we examine representative GGCM-crop-irrigation-region combinations under SSP585 and compute global CO_2-yield responses with and without the inclusion of individual extreme-response regions. The results show that, while these regions account for only a very small fraction of the dataset (~0.02%), they can exert a disproportionate influence on aggregated outcomes. When all extreme-response regions are excluded, the global CO_2-yield response follows a smooth trajectory. In contrast, including one extreme-response region can amplify the global aggregate by up to a factor of ~3 under high [CO_2] (e.g., CYGMA1p74-*swh*-*noirr*-JOR). These examples demonstrate that a small number of regions can alter aggregated responses, particularly at high CO_2 levels.

We emphasize that masking is applied in the main analysis to ensure robustness of aggregated results, while the original parameters are retained to preserve consistency with the source GGCMs and to allow users to explore these effects if desired. The new analysis and corresponding figure have been added to the Supplementary Information Text S1.

Moreover, we clarified that extreme-value regions are considered from both the original data and the emulated responses by adding the following text to the manuscript:

"In addition to extreme values inherited from the original data, the emulated yield response can also exceed 10. We therefore consider extreme-value regions arising from both the original data and the best-fit functions."

4. For the fully irrigated firr case, the manuscript sets the water stress term to 1. However, in reality, irrigation can be constrained by water availability, competition for water resources, and infrastructure limitations, so conditions are not always fully unconstrained. Should firr here be understood strictly as fully irrigated under the ISIMIP experimental design. If so, it would be helpful to emphasize more prominently in the Discussion the caveats for real world applications, especially in regions projected to face future water scarcity.

Thank you for this important comment.

In our study, the term "*firr*" (fully irrigated) follows the ISIMIP experimental design protocol, where it is assumed that irrigation water is unlimited and applied to meet full crop water requirements. This is a modelling convention rather than a reflection of real-world conditions.

We agree that this distinction warrants clearer emphasis, particularly for interpreting results in regions projected to face future water scarcity. We have added a discussion of this caveat in the revised manuscript, as suggested.

The text is added into the Discussion section:

"While our emulator aims to represent real-world conditions, the fully irrigated (*firr*) assumption implies irrigation water is supplied whenever required to prevent soil moisture stress. This simplification does not reflect real-world irrigated croplands, where water resources and infrastructure constrain irrigation (Elliott et al., 2014)."

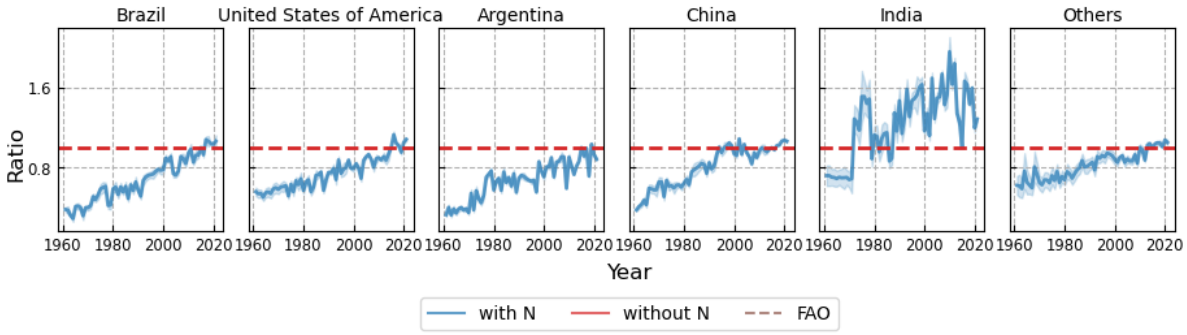
5. The authors explain that GGCM based samples are insufficient to constrain nitrogen responses, so they rely on response functions derived from long term field experiments. They also refit cereal responses using forms such as Michaelis Menten, Mitscherlich, and George, selected using BIC and R squared, noting that limited samples lead to non region specific responses. How do the authors address potential biases arising from this hybrid structure, where climate responses come from GGCMs while nitrogen responses are taken from experimental meta functions.

Thank you for raising this point. The hybrid structure reflects both a practical constraint and a design choice of the emulator. The available GGCM simulations are from GGCMI Phase 2, while the climate responses are derived from GGCMI Phase 3 results. In addition, the nitrogen experiments in GGCMI Phase 2 include only a limited number of nitrogen input levels, which is insufficient to robustly constrain nitrogen response curves. Therefore, we use response functions derived from long-term field experiments to constrain the shape of crop responses to nitrogen inputs.

In this context, combining GGCM-derived climate responses with experimentally derived nitrogen response functions is a strength of the simple emulator framework, as it integrates complementary sources of information. The GGCM ensemble anchors the emulator's baseline yields and climate sensitivities, while the experimental data provide empirically grounded nitrogen response behavior that cannot be robustly inferred from the GGCM samples alone.

6. The manuscript assumes that nitrogen is not a yield determining factor for soybean and therefore sets the N response to a constant. However, in the FAO comparison, soybean yields tend to be overestimated, and the manuscript explicitly mentions this assumption as one possible cause. In which regions does the assumption soy N response equals 1 cause the most pronounced problems. As a minimal alternative, the decision would be far more convincing if the authors tested a simple form such as a weak saturating response with very small sensitivity or a simple upper and lower constraint depending on N input level and compared performance.

Thank you for this helpful suggestion. To assess where the assumption of a constant nitrogen response for soybean has the largest impact, we compared temporal yield trends between the emulator and FAO data by normalizing both time series to their 2015 values and evaluating their ratio. This analysis indicates that the overestimation is the most pronounced in India among top producing regions.



We acknowledge the reviewer's suggestion to test a saturating nitrogen response. However, we did not identify sufficient and consistent empirical evidence to robustly constrain a nitrogen response function for soybean. Given the lack of reliable calibration data, introducing an assumed function would risk adding poorly constrained behavior to the emulator.

Instead, we adopt a conservative approach by assuming a constant nitrogen response for soybean, which is consistent with its nitrogen-fixing characteristics and avoids overparameterization. We consider the integration of soybean nitrogen-yield response as an important direction for future work, as nitrogen fertilizer sensitivity experiments are not yet available from ISIMIP but are expected in the future.

7. The manuscript reports that in sample global correlations are mostly above 0.8, and out of sample differences are generally within plus or minus 0.5 ton per hectare. However, the out of sample test uses different climate forcing while still effectively reproducing the source GGCM simulation space. Do the authors have plans or results for a structural independence test, for example leave one GGCM out calibration and validation, to better assess generalizability.

   Thank you for this suggestion. In this study, we did not perform a structural independence test across GGCMs (e.g., leave-one-GGCM-out validation). Our objective was instead to emulate the collective response space of the GGCM ensemble, and we therefore included as many GGCMs as possible in the calibration to capture the widest possible range of crop model behaviors.

   The out-of-sample evaluation focuses on testing the emulator under different climate forcings while remaining within the response space represented by the GGCM ensemble. In this sense, the emulator is designed to reproduce the ensemble behavior rather than generalize to an unseen GGCM structure. A leave-one-GGCM-out experiment would indeed provide an additional perspective on structural generalizability, and we agree that this would be a useful extension for future work. We have clarified this point in the revised manuscript.

   The added text is in the first paragraph of Section 4.2.

   "Because the emulator is calibrated using simulations from multiple GGCMs simultaneously, it is designed to reproduce the collective response space of the GGCM ensemble rather than the behavior of any individual model. The out-of-sample validation therefore evaluates emulator performance under independent climate forcings while remaining within the range of crop responses represented by the GGCM ensemble."

8. The authors state that they use a 5 year moving average in calibration to filter short term variability, and they acknowledge that extreme event impacts are not considered and that only multi year trends are represented. Could the authors clarify more explicitly the scope of applicability and non recommended uses, particularly what misunderstandings might arise if users apply this model to risk and extremes assessments such as heatwave damage.

   Thank you for this suggestion. The emulator operates at an annual resolution and can therefore capture interannual yield variability driven by year-to-year climate variations. However, because the calibration uses 5-year moving averages, the emulator cannot resolve impacts from within-season extreme events such as heatwaves, droughts, or floods.

   We have clarified in the revised manuscript that the emulator is intended for long-term analysis and scenario exploration, and should not be used to quantify damage from extreme events. This limitation is now stated more explicitly in the Discussion section with the following addition:

   "Moreover, it should be noted that the emulator is suited for applications focusing on long-term climate responses and scenario exploration rather than assessments of extreme events such as droughts and heatwaves (Lesk et al., 2016; Zampieri et al., 2017)."

Minor comments

1. In the Fig. 10 caption, dotted bule lines appears to be a typo and should be blue.

   Thank you for catching this typo. We have corrected "bule" to "blue" in the Fig. 10 caption.