Author response to reviewers for the manuscript egusphere-2025-862 - Folberth et al. "CROMES v1.0: A flexible CROp Model Emulator Suite for climate impact assessment"

We thank the reviewers for their insightful comments. Below we provide a point-by-point response and how we intend to address the comments and suggestions in the revised manuscript. All line number references refer to the revised manuscript with tracked changes using "Simple Markup".

Reviewer 1

Comment 1. This article presents an emulator for global gridded crop models. Although emulators are now commonly used in climate impact assessment studies in lieu of full crop models, the novelty in this article originates from their methodology and not necessarily the approach itself. The authors demonstrate a faster and more efficient pipeline to extract climate features and train the emulators, which yielded an advantage of at least a magnitude over gridded climate models. Although the computational advantage is not explicitly compared to other emulators, the authors sufficiently discuss prior studies and acknowledge the fact that it is not straightforward to do such benchmarking comparisons due to differences in crop models, inputs required, and algorithms being used. The article is well written and is sufficiently detailed to allow for reproducible results with its supplementary code and data, and meets the quality standards for publication.

Response

Thank you for the positive feedback on our manuscript and the study itself. We see from both your comments and the second reviewer's feedback that a key advantage of our emulator suite – the high accuracy of predictions for unseen climate data – may not have fully come across in the original manuscript. We have hence further elaborated on this in the revised manuscript as elaborated in the responses to reviewer 2.

Reviewer 2

Synopsis. The manuscript presents gridded crop growth model emulators for climate change projections analysis. It seems to have a significant improvement in computational time performance without major losses in accuracy. However, there are a few issues that is not clear to me that I would encourage the authors to address prior to publication.

Response

Thank you for the overall positive appraisal of our manuscript and the constructive comments. We have revised the manuscript accordingly and provide below a point-by-point response.

Major comments

Comment 1. To really nail the flow of this work, overall, the paragraphs lack a concluding (so what?) sentence in the end. For example, the first paragraph is missing a closing sentence for the idea that GGCMs are computationally demanding, but there is hope with emulators. Something like: "This high computational cost of GGCM hinders more comprehensive scenario analysis and prevents the quick adoption of new climatic datasets that can be addressed with emulators". You do this in the third paragraph; I would recommend that authors do this in all paragraphs. This is essential in the introduction and discussion sections.

Response

Thank you for pointing to this. We have added a concluding statement along the lines suggested in the comment. The new sentence states (L46):

"*This high computational demand of GGCMs consequently limits the adoption of higher resolution climate forcings or wider sets climate projections that would allow to derive more robust and comprehensive climate impact estimates.*"

We have now also substantially elaborated on the Discussion section in response to your other comments and have taken case to provide a better flow.

Comment 2. There are a few typos, double check with a word processor or something, e.g. Line 44, you mean 1 km, not 1k. Line 157 has 'is' doubled ("PET is (see sect. 2.4.3 for details) is used…"), or in line 201, where the 'i' should be italic.

Response

Thank you for pointing out these typos. We have corrected them (e.g., L44 of the revised manuscript) and screened the remainder of the manuscript, which we expect to be free from spelling errors now.

Comment 3. Clarity is needed for the training and evaluation of the emulators. In Section 2.6, RMSE of yield? So, is it 0.447 t/ha? What is the cut for the climate? Was the 4-fold CV randomly sampled, or were the folds defined based on time or region? The following section (2.7) states that the evaluation is performed across all individual locations and years. I think that means there is data leaking between the training of the emulators and its evaluation. I understand that there are computational limits to this study but it is not clear to me how this was actually done. There is evidence that one can draw a lucky strike on the splits, and that agricultural problems need specific cross-validation strategies depending on the uses (https://doi.org/10.1016/j.compag.2023.107642 & https://doi.org/10.1007/s11119-024-10212-2 & https://doi.org/10.1175/AIES-D-23-0026.1). I would expect models to be trained and the 4-fold CV for hyperparameterisation to be performed with the 1980-2014 data, and evaluation to be performed with the future projections.

Response

Thanks for catching the missing units for RMSE. We have added these [t ha^{-1}] now throughout the manuscript where lacking.

Based on the reviewer's subsequent questions and comments, we see that this comment was in part due to a misunderstanding due to the limited information provided on emulator training and evaluation. This part has been substantially extended now in the manuscript. Essentially, we do not train emulators for a historic time period such as 1980-2014 to apply them to future scenarios nor do we evaluate the emulator performance solely on the same data that have been used for training (or mix training data with test data as would be in the case of training data leakage). As elaborated in response to your comment 6 (and in the revised manuscript), we train emulators on individual GCMs and evaluate the emulators' performance on those same training data as a basic reference metric of how good the algorithm can fit the training data. However, the main evaluation of emulator performance to produce predictions for unseen climate data is done empirically by applying the emulators to climate projections from four other GCMs not used in the training.

This also has implications of how the cross-validation (CV) needs to be designed. In statistical learning problems, the machine-learning framework typical for tabular data, CV is a standard method for estimating models' generalization error with the primary goal to select a model minimizing it. This renders CV design critical for cases of limited samples and highly flexible algorithms, where the risk of overfitting is high (e.g., Vapnik, 1998). However, the training of an emulator for a process-based model as done herein diverges from these common machine-learning paradigms in two key ways:

1) Objective difference: While we use machine-learning methods to build the emulator, we are not primarily concerned with minimizing the error for the training data distribution since the emulator's core purpose is the application to different but still comparable data distributions (here other GCMs). Hence, any marginal gains on the training domain may not have any benefit to the out-of-domain application. Therefore, we opted for hyperparameter grid search with 4-fold CV as an economical approach.
2) Data availability: We generate massive sets of training data by running the process-based model, here EPIC, repeatedly. This virtually eliminates sample size constraints and allows for getting robust estimates for performance metrics.

As lined out above, our training setup uses a fixed GCM with the corresponding three SSPs and the historical period. This yields a tabular dataset with around 45 x 10^6 observations and 60 features. We treat this composite dataset (representing 151k simulation units with 285 years of climate across all climate scenarios) as the distribution of practical interest. In contrast to the references provided by the reviewer dealing with fairly limited training data, this comprehensive sample size does not require any elaborate CV schemes. The actual evaluation then takes place by estimating the accuracy empirically through the application of the emulator to another very comprehensive sample not used in the training and four times the above size (that is, four GCMs x the above four climate scenarios).

We still agree that it is relevant to provide insights on how statistically robust our evaluations presented in Table 3 and 4 of the manuscript are. To do so, we have computed a bootstrap

approximation of 95% confidence intervals for RMSE for both the training set and target sets for each emulator. The RMSE confidence intervals for all datasets are included in a new Table A1 in the new Appendix A. Essentially, they indicate very narrow confidence intervals, supporting our above elaborations.

In line with these points the revised Methods section now states regarding emulator training (L313):

"*With fixed depth = 14 and iterations = 1600, the remaining training parameters were left to default values. For further emulator training, climate scenarios (i.e., historical and three SSPs; Sect. 2.9) were pooled for each GCM separately and emulators trained on the whole sample as the other four GCMs not used in the training were subsequently used as novel data for benchmarking (see subsequent sections). This setup differs from the more common approach of training machine-learning models on historical data with extensive CV and applying them on future scenarios (Richetti et al., 2023; Sweet et al., 2023). Here, models generalize over scenarios rather than time, and similar data distributions and levels of correlation are expected. To support our assumptions, we provide bootstrapped RMSEs with confidence bounds that show the generalization ability of the model (see sect. 2.7.).*"

and with respect to emulator evaluation (L330):

"*In line with earlier studies on crop model emulator development (Blanc, 2017; Franke et al., 2020b; Oyebamiji et al., 2015), we use the root mean square error (RMSE) and linear regression statistics (Pearson's correlation coefficient R2, slope, and intercept) to evaluate emulator performance. The first also corresponds to the metric for the loss function in emulator training (see sect. 2.6). To evaluate the robustness of mean RMSE estimates across the whole sample, we estimate 95% confidence intervals (CI) bootstrapping 500 subsets of 100k samples each. We provide all metrics for two sets of benchmark data:*

(1) We evaluate the performance on the training data itself to show how well the model can fit these training data (sect. 3.1), which also serves as a reference for evaluations on unseen target data.

(2) The main objective of the performance evaluation, however, is the emulators' skill in predicting crop yield simulation outputs for climate projections that have not been used in emulator training (sect. 3.2). Essentially, we train individual emulators for each of the five GCMs used in this experiment (see sect. 2.9) and then apply each of these emulators to the remaining four GCMs not used in each emulator's training. This serves as a vast empirical test of how well the emulators perform on unseen climate features while staying within a comparable domain of climate projections."

The corresponding results state (L412):

"[...] *The 95% confidence interval width for RMSE on the training data is for all GCMs ≤ 0.01 t ha-1 or ≤ 2% of the mean RMSE (Table A1) indicating highly robust results.*"

and (L446):

"[…] *The width of the 95% confidence intervals are with uniformly ≤ 0.3 t ha-1 (Table A1), corresponding to ≤ 5% of the mean, as well marginally higher than for the training data but still very low in both absolute and relative terms.*"

In the Discussion we state now (L599):

"[…] *Rather than a CV, we performed here a bootstrapping of emulator predictions to quantify 95% CIs for RMSE and found robust results for both our training and application of emulators.*"

References

Vapnik, V.N., 1998. Statistical Learning Theory. Wiley.

Comment 4. Looking at the results, it is expected that the ML will have a nearly perfect fit with the training data, and if you're using GCM data to train future scenarios, what is the value of the emulators? You've already run the GCM on that scenario! In this sense, section 3.1 should be supplementary materials.

Response

We provide in section 3.1 the performance of the emulator on the training data as an elementary metric. Otherwise, numbers in the subsequent sections cannot be put in context of these training metrics. In line with your other comments (e.g., comment 7), we have further streamlined the methods sub-sections and the presentation of results to make sure that it's clear where emulators are applied to unseen climate projections and where on training data.

Comment 5. If we were to use CROMES in a scenario the ML never saw, how would that perform? Because then we can say, look CROMES allows us to run x more scenarios in the same time GCMs would take to run one. This performance boost does not reduce the quality of the information.

Response

As you state in your next comment, this is precisely what we do in the subsequent sections of the results, I.e., as there is no theoretical guarantee that an emulator would perform well on an unseen set of climate projections, we do extensive empirical testing on many-to-many GCMxSSP combinations. We have now further elaborated on this in the Methods section in response to your comment 6.

Comment 6. After looking back at the M&M I found one tiny sentence stating that training and evaluating the MLs are in different GCMs. Make this more obvious and clear, having to come back to M&M and go back and forth means this is not clear (or that I'm dumb, which might be true as well, but let's make it easier for the readers).

Response

We agree that the approach to emulator performance evaluation has been scattered across the manuscript and somewhat obfuscated in the Methods section. We have now elaborated on this in sect. 2.7 Emulator evaluation metrics (L330) as cited in the response to your comment 3.

Comment 7. I would recommend having a session. "Training, variable importance, and evaluation" where you clearly state the different experiments to assess the MLs performance. When you say that the evaluation was performed with the GCMs not used in the training, is it a CV scheme where you leave-a-GCM-out or is there a GCMs used for training and a set for evaluation?

Response

As lined out in response to your prior comment, we have now elaborated sect. 2.7 to better communicate the emulator evaluation. We would prefer to keep sect. 2.6 (Emulator training and feature importance) and sect. 2.7 (Emulator evaluation metrics) separate as both should topically be sufficiently well separated. A joint section in turn may become lengthy and thereby more challenging to grasp.

Comment 8. Good that you made a comment on the negative yields and the fragility of R2 with the great amount of points.

Response

Thanks for the confirmation. Yes, these are important points we think need to be mentioned. Indeed, R2 still remains the most common metric in this area of research, rendering it the key statistic for comparison with other studies. While RMSE is the objective function of the emulator training, RMSE is also influenced by absolute yield that can vary greatly among GGCMs providing the emulator training data and hence more challenging to compare among studies.

Comment 9. Ok, I think authors could concentrate on highlighting where CROMES performs the worst with bigger plots version of Figures 4 and 5. Currently, you can barely see any difference. All the rest could be supplementary materials. Authors want to state something like: CROME helps with GCM. The worst case is this and this, here and there, everything else is better (see supplementary materials). This way the important results can be clearly observed and the key points made without losing any of the detail, as is, it is a bit hard to follow what are the key findings of the section as opposed to the Feature Selection where it is much easier to understand that the most important variables for the emulators are x, y, and z. Same for the computational performance.

Response

We agree that a more nuanced presentation of emulator performance is helpful. However, as the emulators do show good performance in the experiment herein, we do not see a reason to focus on cases of poor performance alone. We hence opt to keep the current presentation of emulator performance as a basis while extending it with more information under which broader conditions the emulators perform better or worse. To do so, we have masked the evaluations of pixel-wise R2 by (a) major Koeppen-Geiger regions and (b) extent of harvested area. The respective figures form part of a new Appendix A and the text in the main body has been extended to cover the new results. Essentially, they underpin that (a) R2 tends to be lower in (semi-)arid regions but is not uniformly low there and (b) performance is poorer where harvested areas are small. The new results state (L510):

"*These visual interpretations are supported by density plots of R2 per Koeppen-Geiger region – major global climate domains – for the whole land (Figure A1) or pixels with maize harvested area > 100 ha (Figure A2). The first shows that a comparably high tail occurs in (semi-)arid climates and a flat distribution is found for polar climates. Both present challenging environments for agriculture and in the latter case have hardly harvested areas. Accordingly, when removing pixels with marginal harvested areas, the distributions across all climates shift towards higher R2 values. The higher performance in pixels with larger harvested areas is reinforced by Figure A3 displaying R2 densities within harvested area bins. The highest tail towards low values is again found for pixels with areas < 10ha, whereas pixels with very large harvested areas (> 1000ha) have hardly R2 values of less than 0.5.*"

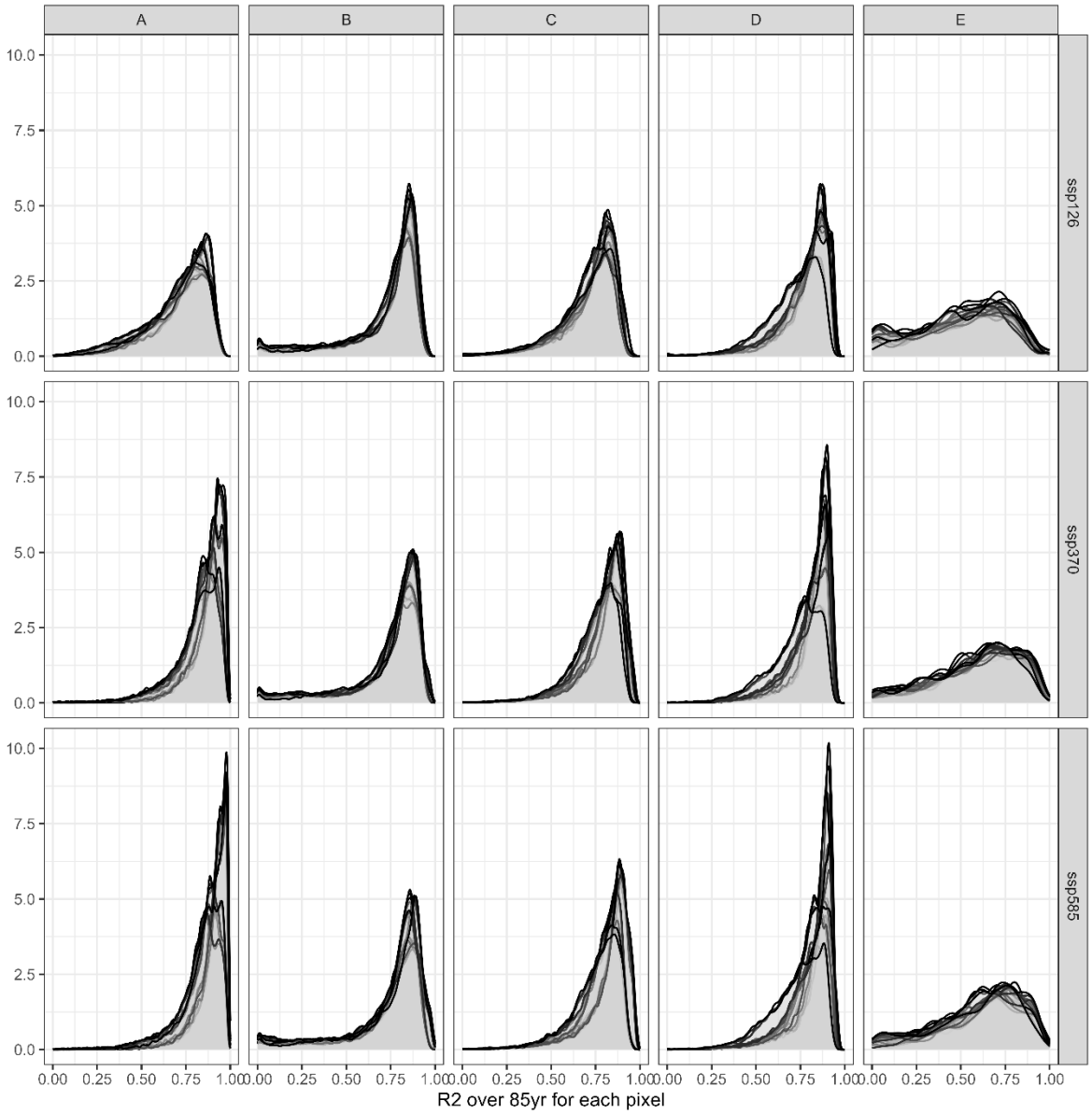


Figure A1. Density of R2 per pixel over 85 years for three SSPs (rows) and across five major Koeppen Geiger climate regions (columns). A=tropical, B=(semi-)arid, C=temperate, D=cold, E=polar. Each panel shows 20 plots, applying each emulator trained on one of the five GCMs to the four GCMs not used in its training.

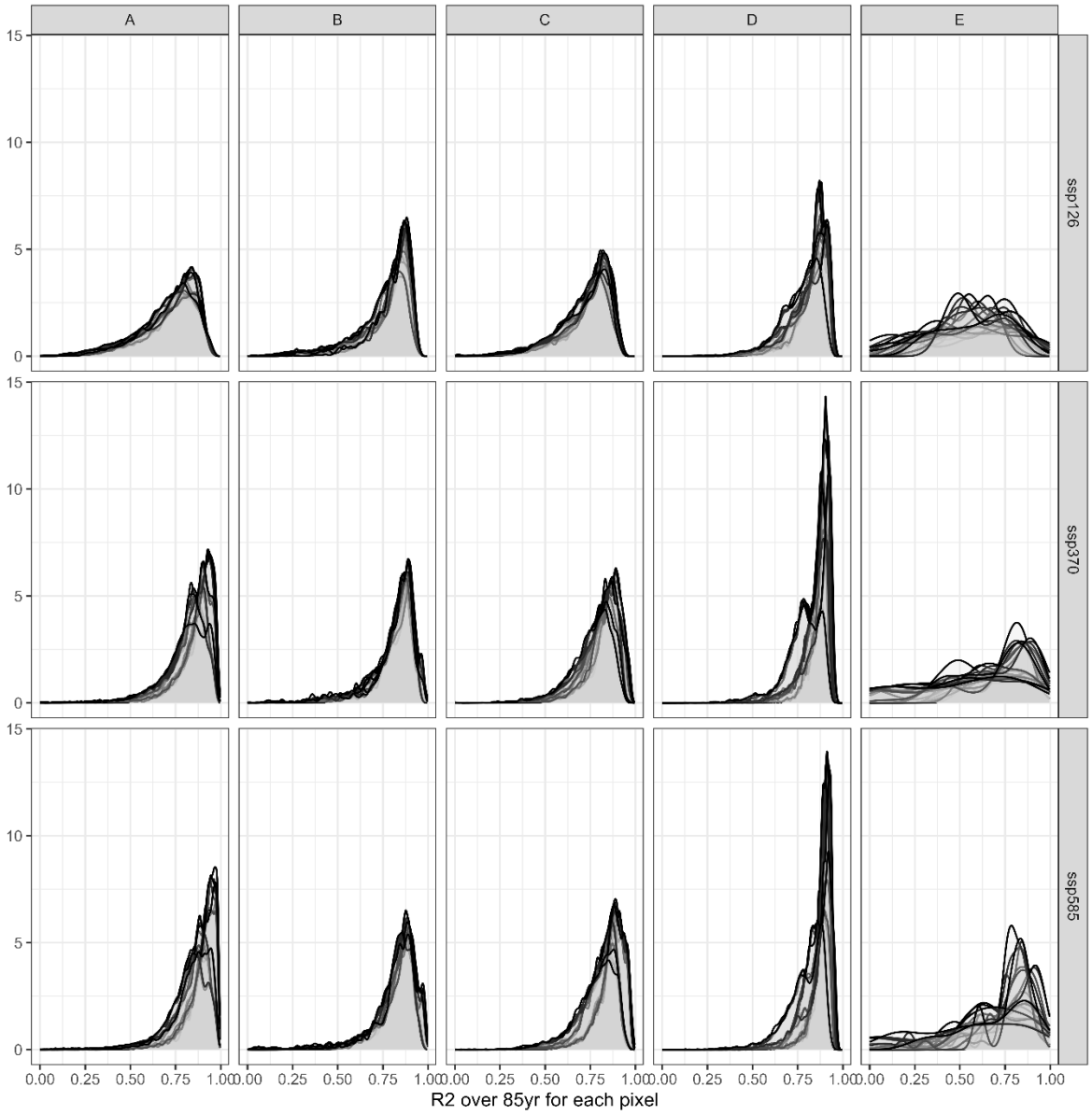


Figure A2. Same as Figure A1 but showing only pixels with rainfed maize harvested area > 100ha.

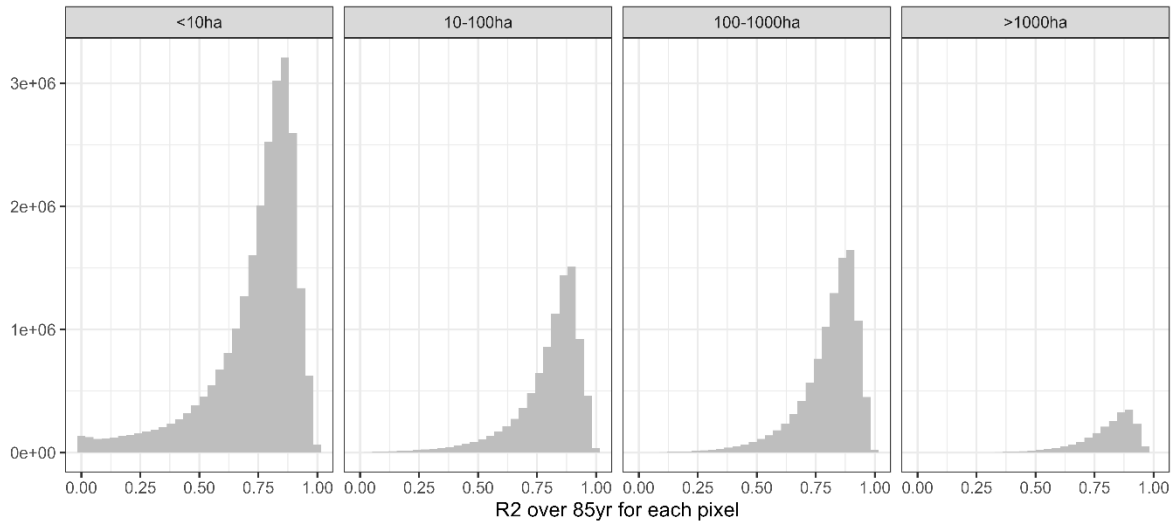


Figure A3. Density of R2 per pixel over 85 years for four bins of rainfed maize harvested areas (panels). Data are pooled from applying each emulator trained on one of the five GCMs to the four GCMs not used in its training across all three SSPs.

Comment 10. Section on computational demand could be one paragraph concentrating on the performance, there are a lot of text already present in the M&M.

Response

In line with responding to your comment 21, we have streamlined the section on computational performance and condensed the first two paragraphs into one (L546). Please see the response further below for a quotation from the manuscript.

Comment 11. The discussion is a bit shallow.

Response

We have extended the discussion in response to your other comments.

Comment 12. First paragraph, concluding that low-yield areas matter less, I would say they might matter more, as the food security of those areas is more compromised! They might be of less economic importance on a global scale, but those are the areas that might suffer more with decreasing food security due to climate change!

Response

This must be a misunderstanding. We don't state that low-yielding regions don't matter but that in arid climates (L578) "*[…] rainfed agriculture is typically of limited importance [in such regions] and*

the constantly low crop yields pose a challenge to achieving a good regression fit while the absolute error can be considered minor."

Still, we agree that a more nuanced discussion should be provided on where the emulators trained in this exemplary experiment perform better or worse. To do so – and in line with other comments -, we discuss now more aspects of emulator performance and training with respect to potential applications (L580):

"*Overall, the performance of an emulator will need to be evaluated on an application case basis and training routines may need to be adjusted for specific target regions or applications to obtain best results for a specific context. For example, where farming in semi-arid environments or other low-yielding regions is in the focus, the selection of training samples should be tailored to such regions to ensure that the algorithm is not geared towards a mean response that covers a variety of climates where semi-arid conditions present a particular niche. Vice versa, when focusing on breadbasket failures, users may sample such typically high-yielding agro-climatic regions specifically. In the demonstration case herein, that is tailored towards evaluation for broader coverage of global climate projections, we selected accordingly all pixels globally.*"

Comment 13. Second paragraph, starting on line 535. So what? What is the conclusion of all the various studies in relation to this? Sweet et al. (2023) state that different CV schemes impact the outcome. What does that means in this study? I think you need to talk about what you have chosen and how that impacts the performance, i.e. why yours look so good?

Response

Thank you for pointing out this shortcoming. In line with responses to your earlier comments and edits to the manuscript such as on CV, we have now extended this part of the discussion. This includes further elaborations on different studies' approaches to climate feature estimation in line with the response to your following comment 14. The revised paragraph states (L590):

"*To the authors' best knowledge, complex machine-learning algorithms have not been applied prior to train emulators for a GGCM using opportunistic training samples, i.e. data that are readily available from earlier experiments. The performance achieved herein is hence not straightforward to compare to that found in earlier studies. Most recently, Sweet et al. (2023) evaluated CV strategies for training machine-learning algorithms to predict crop yields from GGCMs. They reported a maximum R2 of 0.82 on the training set and far lower values around 0.4 on holdout data. However, their application case covered only the historic period and focused on holdout years and regions, which may be more challenging to capture than multi-year and –location climate change projections as herein. Yet, they also assumed static growing season lengths, which does not reflect the conceptualization of plant maturation typical in crop models and loses information on the weather the crop is actually exposed to (see also next paragraph). Rather than a CV, we performed here a bootstrapping of emulator predictions to quantify 95% CIs for RMSE and found robust results for both our training and application of emulators. Oyebamiji et al. (2015) developed a similar emulator approach as the one herein but using various regression methods and with the objective of predicting changes in decadal mean crop yields based on changes in climate features over the*

four meteorological seasons. Applied to an older version of the GGCM LPJmL (Bondeau et al., 2007), they found an agreement with R2=0.72 to 0.86 for unseen climate projections combining RCPs 4.5 and 8.5. Similarly, Blanc (2017) trained statistical emulators for crop yield changes under climate change based on various regression models for several GGCMs and samples from climate impact projections using monthly and meteorological seasonal climate features. This resulted in an R2 of 0.43 to 0.78 for multi-year average yield changes depending on the GGCM with R2 0.48 to 0.56 for an EPIC-based GGCM GEPIC. Finally, Franke et al. (2020b) trained GGCM emulators using pixel-specific polynomials for a range GGCMs that had simulated a structured training sample with systematic changes in temperature, precipitation, CO2, and fertilizer application. Applied to an exemplary climate change projection (HadGEM2-ES with RCP8.5) using annual shifters in climate features this resulted in RMSE of 0.9 to 2.7 t ha-1 and 1.8 to 2.4 t ha-1 for two EPIC-based GGCMs compared to herein R2=0.97 to 0.98 and RMSE=0.50 to 0.66 on holdout data."

Comment 14. Third paragraph, why you start talking about what is not in scope instead of stating what you found and its implications? CROMES incorporates phenology like all the crop models, are these shown as key important variables? What does that means? Why should we care?

Response

Thanks for the suggestion. We have now largely rewritten this paragraph, elaborating on the importance of feature engineering and placing the potential for future systematic evaluation of the importance of feature engineering at the end (L615):

"*We expect that feature engineering is the key determinant for the high accuracy of crop yield predictions, also compared to past research. Earlier studies developing emulators or similar hybrid crop modelling tools employed fixed seasonal, monthly, or annual aggregates of climate variables (Blanc, 2017; Folberth et al., 2019; Franke et al., 2020b; Goulart et al., 2023; Oyebamiji et al., 2015; Sweet et al., 2023). These provide basic information on the weather a crop is exposed to in a specific year but neglect that crop maturity is driven by temperatures, represented as GDD accumulation in the vast majority of (global) crop models (Jägermeyr et al., 2021). Following this concept, CROMES dynamically estimates the actual length of each growing season and its sub-phases based on GDD accumulation. This has earlier been found to be a key determinant of crop yields in GGCMs, especially under high levels of global warming. Essentially, crops mature earlier and have less time for biomass accumulation but may simultaneously not be affected by adverse weather events later in the year (Zabel et al., 2021). A systematic comparison of different feature engineering approaches, however, is beyond the scope of this study and should be subject of a dedicated intercomparison exercise as is common within the crop modelling community for process-based types of models.*"

Comment 15. Last paragraph, ok. Why do I want to make quicker GCM simulations? What is the actual benefit for the community?

Response

The relevance of speed and use cases of the emulator suite have now been further elaborated in the Introduction (see response to your comment 1) and in the Conclusions, incl. the advantage of higher speed in obtaining crop yield estimates. The paragraph now states (L640):

"*We expect the crop model emulator pipeline presented herein to bear great potential in various applications including complex climate impact modelling clusters or comprehensive scenario analyses across large climate ensembles and at high spatial resolutions. For such applications computational efficiency is a key advantage and the loss of information compared to the gain in speed achieved herein indicate that outcomes can be considered robust as long as predictors are part of the training domain. Quantifying this validity domain remains a prevailing issue in machine learning and will have to be characterized on a case-by-case basis until robust methods are developed. This will be an important subject for future research. Meanwhile, compared to static emulators CROMES allows for continuous updating of training data such as for the next generation of CMIP7 climate projections, with new GGCM versions, or for applications with very specific feature domains such as global cooling scenarios from geoengineering or nuclear winter. Thereby, no tailored crop model simulations are required for training as long as data from existing experiments are within the application domain and users of the emulator pipeline do not require specific expertise in crop model setups and applications.*"

Minor comments

Comment 16. Line 145: I would expect times for this in the results

Response

The numbers provided here are solely semi-quantitative estimates to highlight the relevance of the initial netCDF to binary conversion as part of the whole computational pipeline, based on experience from earlier software engineering. We have not tested different types of data access in this study and can therefore not provide further quantifications of how I/O speed would be affected in case of non-optimized access methods. However, we see that further explanations would be helpful here and have further elaborated the paragraph (L141):

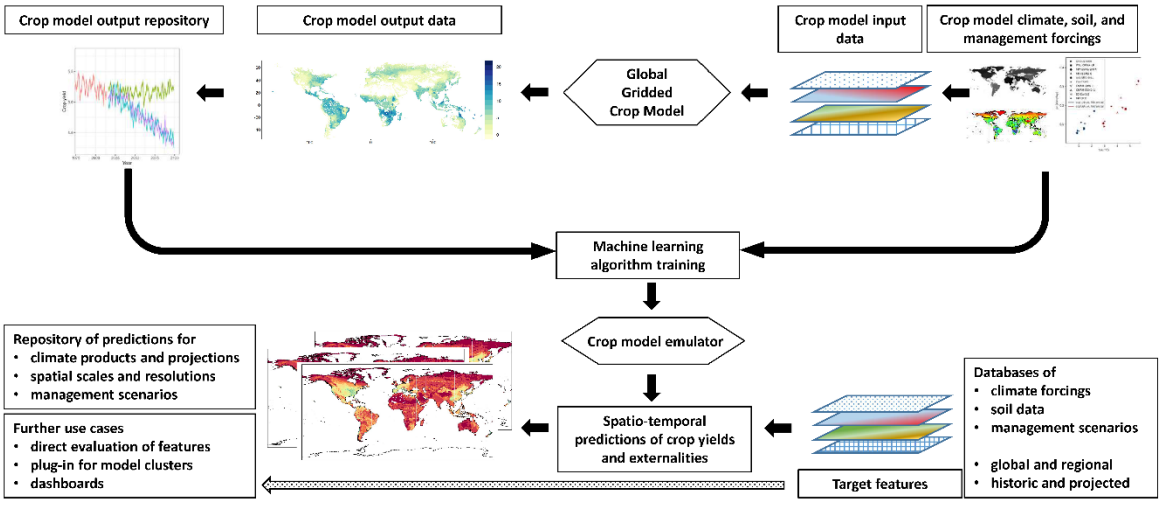
"*The conversion carried out once per climate data set substantially speeds up the subsequent climate feature engineering process. Selecting all climatic values sequentially for each individual map pixel is infeasible due to the large size of the pixel set (here, the ISIMIP 3b cropland mask with 65797 pixels) and the large number of days (about 36500 for a 100-year dataset). Together with the number of climatic variables (here six) this leads to about 66000 * 36500 * 6 = 14 * 109 selection operations from individual files. As one selection (seek) operation on a state-of-the-art solid-state drive can take more than 0.01 to 0.2 ms, this would result in 14 * 109 * 0.01 / 1000 / 3600 / 24 = 2 to 40 days of processing, assuming that data is not loaded into computer's memory or cached. This*

*bottleneck can be solved in a straightforward manner, if there is sufficient memory available on a user's computer, but the memory consumption would be close to 360 * 720 * 36500 * 6 * (4 bytes/value) = 210 GB for loading all uncompressed netCDF files into memory. To substantially speed up climate feature processing while avoiding large memory requirements, our implementation carries out a data format conversion through a dedicated routine that is extensively using a small portion of RAM (less than 1 GB) by handling netCDF files individually and producing intermediary binary files. These can subsequently be used for sequential data processing that avoids intensive seek operations or extensive memory use. This allows to (1) reduce running time down to few minutes, (2) avoid dependence on high-end hardware, and (3) supports parallel runs in a high-performance computing environment."*

Comment 17. Figure 1. Make the text in the boxes bigger. There's no point having them if they're it's hard or impossible to read.

Response

We have increased the font size, switched to bold formatting, and removed redundant visuals to improve readability.



Comment 18. Line 219: from seed/planting to emergence is 100oC days? 100GDD growing degree-days?

Response

Based on equations 1 and 2, the unit [°C] is technically correct for GDD although it's common jargon to use GDD or HU as a unit identifier. To make it more specific, we have rephrased to (L235):

"Prior to emergence of the crop, an additional amount of germination HU (GMHU) is required for the seed to develop to a seedling, here a GDD sum of 100 °C for maize."

Comment 19. Line 224: What does the cut-off mean? 21 days after planting, the maize will mature? I would expect, even for a short-season maize, to take at least 30 days to reach reproductive stages. Looking at Figure 2, the cut-off is not after planting. Please clarify this.

Response

Thank you for catching this. Indeed, it refers to the reported harvest not planting date. We have corrected this in the revised manuscript and the sentence states no (L241):

"*If the crop does not mature due to too low growing season temperatures, a cut-off is enforced 21d after the reported harvest date.*"

Comment 20. Line 345: How much is sufficient N? This is needed for replicability of the study, as you highlighted, it is different from Jägermeyr et al. (2021).

Response

We have now elaborated the parameter settings to allow for better reproducibility of the study. The revised paragraph states (L375):

"*The setup and parameterization of the EPIC-IIASA GGCM was kept the same as in ISIMIP3b (Jägermeyr et al., 2021) except that we used here sufficient nitrogen (N) fertilizer inputs to focus on climate signals. Following this approach, N is applied automatically by the model as required by the crop to meet its demand for biomass accumulation. The model's application threshold parameter BFT0 was set to 0.99, corresponding to N application if N stress limits crop growth by more than 1% compared to the potential, the maximum annual input FMX was set to 999 kg N ha-1 yr-1 to ensure that no N stress occurs.*"

Comment 21. Line 496: you bother to state what GPU stands for here, but not on the crazy acronyms of the GCMs. I think it is more likely for an average reader to be aware of what a GPU is than what UKESM1-0-LL is. Further, most of this paragraph is redundant.

Response

We agree that the readership of GMD should be familiar with the acronym GPU. We have removed it in the revised manuscript and leave it to the editorial office whether to include the spelled-out name in the final version. We have furthermore condensed the paragraph and merged with the subsequent one, which now states (L546):

"*As time gain is a key advantage of emulators, we provide a rough estimate of time required for key tasks within the modelling and data processing chains of both approaches - EPIC simulations and emulator training and predictions - to allow for basic contextualization (Figure 7), while actual performance in individual applications will depend on the computational infrastructure in place and its load. In the setup used herein, both approaches require first a conversion of netCDF files to binary files that provide substantially faster read access. This takes about 0.5h. Further production of daily weather files for the EPIC model – individual text files for each pixel - takes approx. 2h. The*

largest time requirement occurs for the EPIC simulation itself, which here takes 12h but can vary on the shared cluster between 6h and 18h on a single core. The crop model produces single output files for each simulation unit from which the extraction of outputs to a compilation file requires 1h. Once a climate dataset has been processed, only the last two steps crop model run and post-processing are required for each simulation."

Comment 22. Figure 7: Being pedantic, the colour scheme could be more harmonious. Warm tones for EPI and cold for CROMES? Add a bit of space between CROMES and EPIC in the tasks and the netCDF to binary conversion, which is the same for both.

Response

Thanks for the suggestions, we agree that this would improve readability and have reformatted the figure accordingly.

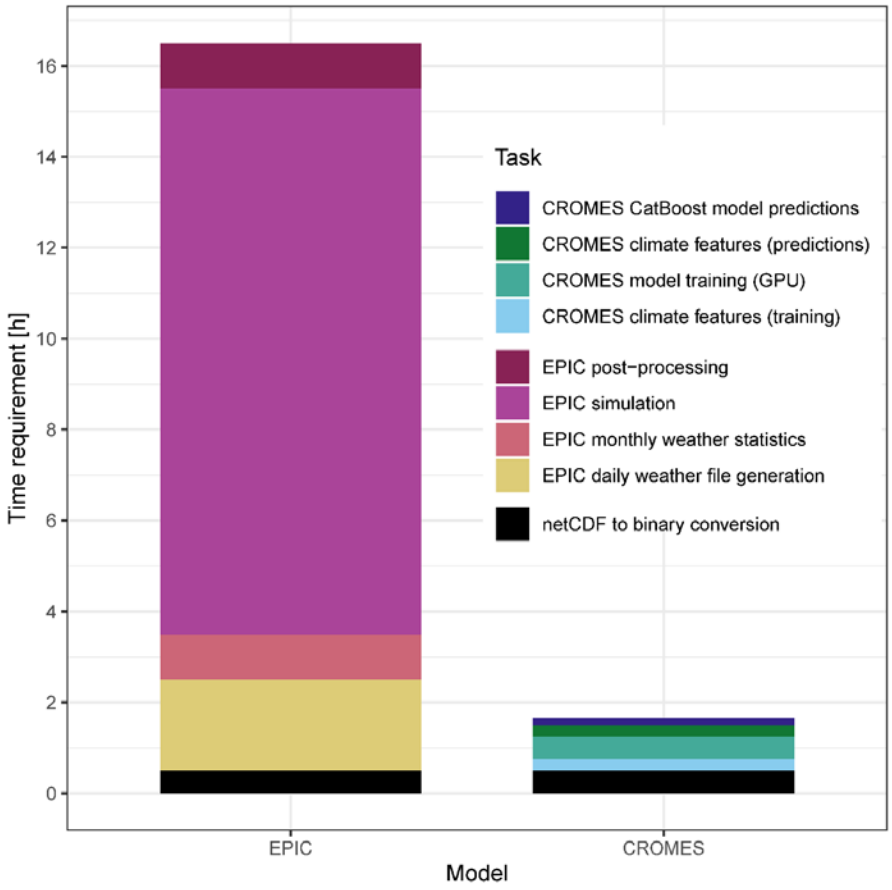


Figure 7. Time requirement for key tasks required to produce global crop model simulations with EPIC or crop yield predictions with CROMES. Some tasks only have to be performed once, essentially the bottom three of the legend or those relating to CROMES emulator training, depending on the specific purpose. The numbers shown here are therefore primarily for illustrative purposes.