Author response to comments on gmd-2019-365

Referee comments in blue | Author response in black | MS change in red (line numbers refer to those in the attached latex diff document)

---

## Referee 1:

1. The study is a step forward in crop model emulation and is in principle a useful contribution to the literature. The paper contains a lot of excellent technical work. I focus here on areas for improvement, which I describe as major simply because I think a re-framing is needed in order to ensure that the paper is used well, and is not mis-used in the future.

Thank you for the assessment. We have added text in accordance with suggestions below.

2. The uses stated in the abstract for the emulators are: "providing a tool that can facilitate model comparison, diagnosis of interacting factors affecting yields, and integrated assessment of climate impacts." It would be good to understand more from the paper about how these different usages are envisaged. In particular, the suggestion that the emulator might be used for integrated assessment lacks evidence. It is far from clear that this would be a sensible step to take, because study is subject to a number of important limitations. Whilst the authors are cognisant of these limitations, not enough attention is paid to them in the way that the work is framed and interpreted.

We have added text in line with these suggestions; see responses below.

3. One limitation is the use of mean yields. "We emulate the climatological mean response, because that is the response of interest in assessments of climate change impacts. . .. Emulation then becomes relatively straightforward, since changes in time- averaged yields are also considerably smoother than those in year-to-year yield response." – L108. Why is mean yield the response of interest? Perhaps it is because it is relatively straightforward, rather than because it is useful per se. Climate variation explains a third of global crop yield variability – Ray et al. (2015) Nature communications. Why do the authors think that mean yields are interesting? There would need to be a clear rationale in the paper.

We believe that changes in multi-annual averages are actually the most useful measure of future interest. While changes to year-to-year variability in crop yields would be important to farmers if they change significantly, the shift that is most relevant to overall economic impacts, and to decisions on choice of crops and planting locations, is that in mean yields. Many climate change impact assessments therefore focus on multi-annual means as the central metric for climate change impacts. In economic assessments that use crop model outputs to inform IAMs or agro-economic land-use models, crop model outputs are also typically aggregated to multi-annual means (Nelson et al. 2014, Wiebe et al. 2015), because land-use changes (in terms of expanding or abandoning cropland) are driven not by short-term (year-to-year) yield variability, but by changes in average conditions. Finally, it is clear that year-to-year variability in yields is only loosely related to mean growing season temperatures, which are the dominant changes in the underlying dataset. In process-based crop models, changes in mean yields are tightly

related to mean temperatures, so we can provide a reliable emulator of these changes. We therefore have focused our efforts on prediction of changes in mean yields.

The reviewer's comment shows that we have not adequately discussed these considerations, and so we have added text along these lines. We have also clarified that the application of our emulators is constrained to research questions in which long-term dynamics are the relevant feature and that short-term dynamics need to work with other tools. We have also added in our concluding sections more discussion of what would be needed to analyze and emulate changes in crop yield variability, for those working with a focus on these time-scales (e.g. Schewe et al. 2017).

Additional text pertaining to the choice of mean yields has been added to Section 2.2, Lines 118 - 125.

Additional text addressing the issue of Ray et al added to line 400.

4. Assuming a rationale exists for assessing mean yields, over what lead times might the emulators be usefully used? As the authors point out, the emulators cannot be used out of sample, thus implying relatively short lead times, before climate changes significantly. However, over the next couple of decades, changes in mean yields are unlikely to be important relative to extremes.

We agree that year-to-year variability is likely more interesting over the next decade (or two). The emulator is designed to provide projections at the decadal or multidecadal timescale to the end of the century (following in line with the RCP-IPCC framework). While it is true that some areas of the globe will exceed 6 degrees at the high end of climate change (e.g. RCP8.5) by the end of the century, this is not the case for many regions or scenarios with lower radiative forcing, especially when considering changes in multi-annual averages. The projection to the end of the 21st century with assumed fixed management, such as the growing season is unrealistic anyway and needs to be interpreted with care (Minoli et al. 2019, Iizumi et al. 2019).

We have added additional language to the discussion to clarify this timescale of interest, the problems with extrapolation, and the limitations with the fixed growing season.

Additional text pertaining to the choice of mean yields has been added to Section 2.2, Lines 118 - 125.

Additional cautions about extrapolation have been added to Section 6, lines 544-547.

Additional notes about the fixed growing season have been added to Section 6, lines 551-553.

5. Assuming a focus on mean yields can be justified for an appropriate lead time, there remains the question of why an emulator is a valid method to use. Two issues need to be addressed here:
6. i. Whether or not the emulator is fit for purpose. Does it reproduce observed yields well? The link to observed yields is tenuous. Error (which should actually be termed "deviation" – since it is not a true error) is defined relative to yields simulated by the underlying crop models. If the emulators are to be used, then one would need to be sure it captures real historical climate impacts. The language on this is imprecise in many places. For example, in the abstract: "... suggesting that effects of changes in temperature and precipitation distributions are small relative to those of changing means." This statement is true only of model space; indeed, it is untrue of observations as Ray et al. (2015) and others have pointed out.

The purpose of the emulators is to reproduce the output of the process-based models, to provide a lightweight substitute for the computationally expensive calculations. We therefore do not focus on validation of those process models in this paper. Extensive model validation exercises were carried out as part of GGCMI Phase I *(*Müller et al. 2017*),* and we address model validation of GGCMI Phase II in the "experiment description" GMD paper (Franke et al 2020a). The emulator is therefore fit for purpose if it captures the output of the process-based models, which we cover here extensively in Section 4.1.

We agree absolutely that variability in growing-season conditions is critical for year-over-year yield variations. This was shown with historical yields in the Ray 2015 study, and is also true for the process-based models here (see Franke et al 2020a). However, we are unaware of any studies showing that **changes** in variability under climate change are important compared to changes in climate means. Capturing the effects of changing variability in climate projections would be problematic in any case because climate models show very little agreement about future changes in variability (compared to their agreement in the change in means), and often struggle to represent historical variability.

Our statement about the ability of our emulators to capture mean yields in a process-based model under a climate model projection, inclusive of any variability changes, is a demonstrably true statement for the GGCMI simulations. It remains an interesting question whether the process-based models are less sensitive to potential future changes in temperature and precipitation distributions than are real-world crops. Some suggestions along these lines were made by Müller et al 2017, which is cited in the manuscript, but we have now clarified the finding.

In general, these comments suggest that we have inadequately discussed the underlying differences between the process-based models used in GGCMI and statistical models based on historical crop yields. We have therefore now better emphasized these points in the manuscript.

Additional notes about the application of the emulator have been added to Section 6, lines 495-530.

7. ii. Is there a better method? Statistical regressions would by definition capture to some extent observed yield responses to weather and climate. The resulting emulators [are] lightweight, computationally tractable " – but so are statistical models. Reasons to use an emulator over a statistical model are presented in the introduction. However, neither the lack of observed yields in calculating skill, nor the lack of model calibration (another limitation; see below), are brought into this discussion. Similarly, what does the focus on yield changes, rather than yields per se, mean for the robustness of the methodology?

Statistical models are being developed by many different research groups and consist of a separate and somewhat distinct approach to process-based modeling. Statistical models have the obvious problem of little data in many geographic regions and no data under future climate change that has not yet happened. Statistical models can only be evaluated on 'held-out' historical data. It is unclear (and perhaps impossible to test) whether statistical models or process-based models are better for future projections.

The emulator is a statistical model, only it is trained on simulated data instead of 'real' data. It has the obvious advantage of leveraging the body of science behind crop models to provide 'data' where none exist in the real world, both in space (where crops are not currently grown) and time (under climate change that hasn't happened yet). Better forms of emulation may be possible within the GGCMI phase II

framework. We hope the simulation output dataset can become a test-bed for investigating different statistical functional forms.

While we do fit an intercept (historical mean yield), the emulator is intended to be coupled with a dataset of actual yields since models are uncalibrated. We therefore stand by the focus on yield change as a better use of the emulator for impact assessment. We feel this is a more robust application of the emulator.

We have added some additional text to the text to discuss these issues.

Additional notes about calibration added to lines 111-113.

8. The other option discussed briefly in the introduce is the use of process based models. The full set of GGCMI simulations is available; surely the emulators are not expected to outperform their masters? Presumably the “lightweight” approach is deemed to be an advantage for integrated assessment. If this is so then the advantage should be clearly presented.

Correct, the emulator cannot be better than the model it is trained on.The advantage of the emulators is that by providing an analytical form for yield based on climate and nutrients, they allow simulating yields quickly under arbitrary climate forcing scenarios, as would be needed in a study of optimal policies addressing climate change, or in an assessment exercise using non-standard climate projections. Even large sets of pre-computed crop model outputs lack the flexibility to be adjusted to the applications’ needs.

The GGCMI Phase II simulations would not typically be used in assessments directly, since they consist of non-physical combination of parameters and non-physical spatial distributions in climate changes. No single simulation represents a plausible future world, but in combination they allow production of an emulator that can capture yield response under many plausible future scenarios.

We have added some additional discussion on this topic.

Additional notes have been added to Section 6, line 528 - 531.

9. The major revisions needed for the paper will follow on naturally from framing it more clearly to demonstrate the uses the emulators can be put to. As is no doubt clear, I think that the rationale for their use in integrated assessment is extremely difficult to demonstrate; but perhaps I am wrong. It would be worth thinking about the conditions (data availability, crop knowledge, model skill, input data availability, ..) under which the emulators might be a preferred option.

We have added text discussing their use in Integrated Assessment Models (IAMs). As described above, an emulation of climatological mean yield is the appropriate input for IAMs and other economics-based land-use models whose land-use dynamics are always, to our knowledge, based on multi-annual mean yields. It would in fact be incoherent for an IAM to make decisions about land-use changes based on yearly yields, since most IAMs we are aware of utilize climatological mean temperature changes as their climate inputs (sometimes only the global average). As mentioned above, mean temperature change is closely related to mean yield change but only very loosely related to yearly yield variations.

The emulators presented here are developed in collaboration with IAM modelers to meet their needs; please note that the co-author list includes IAM modelers. We recognize though that the submitted

manuscript did not sufficiently emphasize the expected end uses, and so have now worked with our co-authors to add new text describing the several projects currently in development integrating these emulators into IAMs.

Text added to section 2.2, line 119 and section 6, line 533 - 535 to address IAM integration.

10. Model comparison and diagnosis are easier to justify – but even here some work is needed to explain how the emulators could be used. The emulators could be used to highlight areas of CTWN-A where there is consensus and where there is not, thus providing clear evidence of where model improvement, and associated observational datasets, are needed.

Indeed, model comparison and diagnosis is one of the primary intended applications.Several publications are currently in preparation that use the emulators described here for just these purposes: studies that diagnose differences in model responses to particular climate and management inputs, or clarify the interactions between parameters (e.g temperature and precipitation, or temperature and nitrogen addition). These studies are not possible using statistical models fit on historical yields, but require process models run over systematic parameter sweeps. We had discussed this in the Introduction, but as the paper is long we realize that it requires additional text in the Conclusions/Discussion describing these studies, and have added this.

Text added to Section 6, lines 526, 531, 560.

11. Methodology is not clearly separated from results More information on the skill of the models that go into emulators would aid rationale. Some models are more skilful than others. Do you expect the MME to be the most skilful simulation? If different models perform better in different regions, why not use this information in the emulators?

The skill of the underlying crop models is described and discussed in the companion paper (Franke et al. 2020a) and in earlier efforts to describe the crop models' skill (Müller et al. 2017). The paper under review is intended as the "model description" paper describing the development of emulators, not a documentation of the process models themselves. The question of if and under which conditions the MME is the most skilful simulation is a question about the process models themselves. This paper focuses on validating the emulators, i.e. on showing that a simple functional form can capture the response of those process models. The emulators can then become a tool for answering exactly the question that the reviewer poses, and we appreciate the suggestion. Note that the emulators are designed for each crop model individually and can be combined and aggregated at the users' choice and needs for specific applications. We have now added text suggesting that emulators can be used to examine regional model performance.

Text added to Section 6, lines 560.

12. Similarly, which processes are included vs not included in the underlying models. How good at threshold responses are these models? Cf "In general, emulator performance is poor anywhere that models show steep yield changes once some threshold has been reached, whether these are abrupt gains or complete crop failures" - I find these cases very important especially when looking at the end of the century.

Indeed. These cases are important and the provision and publication of the emulators, that are described here, allows for making these analyses. Again, this paper is a model description paper, not the final

application of the emulators that could answer all questions that could be addressed by using the emulators. As discussed above and in the paper, one intended purpose of the emulators is to scrutinize model dynamics and identify options for model improvement (of the process-based crop models, not the emulators).

The temperature response at the 30-year mean scale is very smooth in all but a few cases. Discontinuities (steep changes) in yield are more common when some models show no yield under present conditions and then transition to moderate yield under a certain amount of warming. While some thresholds may exist on the high end for temperature at the yearly scale, there are vanishingly few cases where the 30-year mean yield drops to zero under warming.

We have added some text pointing out some of these cases to clarify the point.

Text added to section 4.2, line 346-347.

13. Why different numbers of perturbations used across different models?

The complete set of simulations is computationally very demanding, and so modeling groups were offered a set of participation “tiers” involving different number of simulations. The protocol is described in detail in the companion “experiment description” paper (Franke et al. 2020a). We have added text here to point the reader to this documentation.

Text added to section 2.1, line 112.

14. Use of normalised “error” (should be “deviation” or similar) makes differences between models hard to see and makes results appear perhaps better than they are.

We agree that the normalised error does not provide complete information, but it is a useful metric in the context of multi-model emulation, because it normalizes the errors in those regions where models disagree quite strongly anyway. Put another way, this metric emphasizes the need for faithful emulation of model output in those places where the models best agree.

Note that we have included in the paper a separate metric that is not normalized across models: the “out-of-sample evaluation”. This test treats all models equally and as separate entities, and was included specifically to provide the kind of assessment the reviewer seeks.

We have added language better clarifying the differences between the two separate evaluation methods and now clarify the difference between ‘errors’ and ‘deviations’.

Text added to section 4, line 274.

15. Be clear which data were used for calibration vs evaluation

We assume that ‘calibration’ here refers to the emulator out-of-sample evaluation process. For out-of-sample validation, we use a 3-fold cross validation procedure where 90% of the data are used to train the emulator (calibration) and the held out 10% is used to evaluate. This is repeated two more times and the results are averaged. The exact simulation cases in each fold vary by model depending on which were provided, and would be quite exhaustive to list in detail. For example for a single model, this would

consist of three lists of 675 conditions that were included in training and 75 conditions that were evaluated against. We do not think such a table of 2200 different listed conditions would be very illustrative.

We have updated the text to make the cross-validation procedure more clear.

Text added to clarify the procedure section 4.2, lines 373-379. There was an error in our original response letter. The cross validation is 90% - 10%, not 2/3 - 1/3 as originally stated. The manuscript was correct.

16. "Emulator performance is generally good relative to model spread in areas where crops are currently cultivated and in temperate zones in general" - probably not hard giving that the crop models are not calibrated. I think the whole study should have been done with calibrated crop models.

As with so many things, there are pros and cons with calibration, especially if no suitable calibration target is available. Calibration would be needed if the intent of the exercise were to produce absolute yields. However, we are focused here on understanding model responses to different climate and management inputs, and in forecasts of fractional changes. We feel that those are adequately and perhaps better addressed with uncalibrated models. In the previous Phase of GGCMI (Elliott et al. 2015, Müller et al. 2017), the harmonization of management conditions appeared to lead to very different model behavior in some models. Note also that global-scale crop model calibration poses tremendous challenges given the lack of calibration targets (see e.g. Müller et al. 2017).

The lack of calibration may make our 'normalised error' metric less stringent than it might be, since calibration would likely (but not necessarily) reduce the spread between models. (See e.g. Müller et al. 2017 for discussion of the effects on future projections of calibration to present-day yields.) But, the normalised error is only one means of assessing emulators, and we conduct the second, non-normalised 'out of sample' validation exercise to provide an assessment independent of the inter-model spread.

We have expanded the text on the rationale for using uncalibrated models, and implications for the application and interpretation of the emulators.

Text added to section 2, lines 111- 113.

17. line 115 - put info for figure in caption as not helpful in main text. And in Fig.1 , cannot see advertised labels of a, b, c, d (although perhaps journal adds these later).

Modified as suggested.

Text removed, as suggested, from lines 129- 131.

18. line 195 onwards - would model features that are able to be dropped be the same if the procedure was repeated including non-cultivated land? i.e. in marginal areas, are different factors important for determining yields? This is mentioned below (line 223).

The suggestion of doing feature selection over non-cultivated land was not tested in this study, but it is an interesting question and could be pursued in follow-up studies, since we provide the full and the reduced form of the emulators. We hope that others will extend on the work shown here.

We have added this point to the Discussion.

Text added to line 560.

---

## Referee 2:

Overview:

1. Understanding crop yield response to environmental changes is crucial for food security. Statistical crop models are easier for calculation but the projection capability is constrained by the range of current conditions. The process based crop models aim to capture the yield response to different environmental changes but computational expensive compared to statistical crop models. This study developed statistical emulators for 9 process based crop models using GGCMI phase II simulations. The author well validated the statistical emulators and discussed the caveats and the potential usage, such as provide an alternate approach for impact assessment. The manuscript is generally well written and I only have several minor comments on the method and results.

Thank you for the assessment.

Minor comments:

2. The whole section 2.2 discussed why there are differences in climatological and year-to-year response. This part is very interesting but somehow could divert the readers who are eager to know how the study uses the training data to develop emulators. It could be a better flow if put the section into the discussion section or supplementary.

We know the paper is very long, but we felt this section was necessary here to explain the rationale for developing our emulators at the climatological mean level. This is a key feature of the study and is a point of confusion for other reviewers and readers. We hoped that by separating this discussion into its own sub-section, readers would feel free to skip it if they do not feel that the choice of the climatological mean yield requires justification. We have tried to add a little more structure to the introduction to allow readers to better pick and choose which sections to focus on, and following another reviewer's suggestion, have now tried to better recap the main points of the paper in the Discussion.

Text added to section 1, lines 83 - 86.

Recap added to section 6, lines 496 - 527.

3. The authors need to refine the section 3.1 to give more information on Y and regressors (what temporal and spatial scale). Line 161 mentioned that "Emulating at the grid cell level". So I think equation 1 was fitting at grid cell level. My understanding is that Y is a vector of 30-year averaged crop yields across different uniform changes scenarios (a total of 756 scenarios?) in one GGCMI model. There are 34 terms in equation 1, aren't there over fitting problems when you have a small number of Y (some models did not done all required scenarios) but a large number of regressors. Please comment.

Indeed, the equation was fitted at grid cell level. Overfitting can be a problem, and some models could not be emulated if they provided too few simulations to the GGCMI Phase 2 simulation data set. We felt that the number of simulations provided was sufficiently important that we repeat Table 3 from the companion paper Franke et al. 2020a that describes the GGCMI Phase 2 experimental protocol.

In the best case, the training domain consists of 756 elements in Y, which is more than sufficient for fitting 34 parameters, according to a “one in ten rule”. Not all models have provided the full sample, but we use a Bayesian regularization scheme (that probabilistically weights parameters towards zero) that mitigates overfitting in the cases with fewer samples. The out-of-sample validation is our test of whether overfitting is a problem - we show that the emulators fit with the Bayesian scheme can predict yields not included in the training set even in the model cases with lower sampling, but that overfitting would be a problem with standard OLS. We have expanded the text in this section to better explain the overfitting concerns and why we think they are addressed.

Text added to section 3.1, lines 185 - 189.

4. Line 138: “in the the”. Double the here.

Thanks for the catch. Removed.

5. Equation 2. Some terms are gray in equation 2. Are those the dropped terms? If so, just delete them.

Terms in gray here are dropped. We left them for clarity of comparison. We have added some language to make this more clear.

Text added to line 215.

6. Figure 10 caption. “the five GCCMI Phase II crops”, the authors used this terms several times, but this sounds like there are five special crops that was created by GGCMI Phase II. I think just say five crops is fine. They are common crops. And how many individual models are incorporated here? I guess it is nine. But there are not nine color lines, is that because some lines are underneath the black thick line? If so, please mention that.

We will modify the language to remove the GGCMI designation as suggested.

All models are included in this figure, but not all models provided simulations for all crops and not all models provided simulations across the nitrogen dimension, so the number of lines is less than 9 in some cases. We now state this explicitly in the figure caption.

Text modified in the Figure 10 caption.

7. Figure 11. In the figure legend, the uniform T sounds like each process model was forced with global uniform T. But I think it means the uniform increase of T, uniform DT is better.

Agreed, this is an excellent point. Modified as suggested.

‘Delta’ added to figure 11 caption as suggested.

8. Figure 11. In the caption, “Circles are emulated yearly global production changes”, those are dots, not circles.

Agreed. Modified as suggested.

“Circles” changed to “dots” in figure 11 caption and where referenced in the text.

9. Figure 11. Why there are no open squares on plot b? And in plot c, open squares for 2 and 4 increasing of T is missing. All the three plots showed emulated uniform T lines, why not show emulated uniform T+W for plot b, and emulated uniform T+W+C for plot c?

To clarify: the open squares are not emulations, they are the actual simulation output. The emulated responses are the solid dots.

Note that this figure does not involve process model simulations of yield under future climate projections. Instead, it shows *emulations* of yields under climate projections, and compares these emulated yields to the uniform-offset simulations of the GGCMI phase II dataset.

In the case where only temperature is allowed to change, we can show a simulation that is a direct analogue for an emulation of a climate projection. In the T and W case, both temperature and precipitation are changing in the climate projection, and we have no equivalent uniform-offset crop simulations. We cannot match the simultaneous values of T and W changes.)

We recognize that this figure is complex and the caption is not as clear as it could be. We have adjusted the language to try to better explain what is being shown.

Text added to Figure 11 caption.

In the SI:

10. Page 2. First line “is not uniform tn the GGCMI Phase II”, what is tn? Should be in?

Should be ‘in’. Corrected.

Correction made in supplement.

11. Figure S6: there are no gray lines (Ontario), why? I want to know if Ontario has the same failure in A1 as in A0.

Good suggestion; the requested line has been added to the figure.

PROMET (ontario) line added to Figure S6 in the supplement.

12. Figure S21: The simulated RCP8.5 (open triangle ) were not found on the graph.

This was in error. Caption modified.

Legend modified in Figure S21 in the supplement.

## Referee 3:

1. The authors present a highly detailed description and evaluation of newly-developed statistical emulators for global gridded crop model simulations (as being contributed to the GGCMI Phase II), specifically targeting emulation of mean yield changes due to changed climate conditions. The authors construct these emulators by varying over carbon dioxide concentrations, temperature, water, and nitrogen inputs, and also test the effects of adaptation. In general, this paper is highly useful contribution to the emerging work of global grid- ded crop modeling primarily due to providing a very well tested, relatively low-error, computationally economical, and low data-input means of reproducing and/or running GGCM experiments (again, as related to GGCMI Phase II). Given the computational expense and large data requirements of the GGCMs, it is worthwhile to have an option to run climate-crop experiments with comparatively less “overhead” and relatively high confidence that the emulators overall faithfully represent specific model and (thus en- semble?) sensitivities. I also think that the authors generally did well to note some key uncertainties both in the GGCMs and how these influence the emulators, although a couple of aspects could be addressed a bit more (and I note these below). Ultimately, one of the key strengths of this work is to provide a comparable and easier means of representing geospatial crop responses relative to the GGCMs (which certainly have other uses as full process-models). Thus, with a few minor revisions, this paper makes an interesting and useful contribution to the field, and I anticipate these emulators being put to good use by many researchers exploring mean climate-crop interactions.

Thank you for the assessment.

I do have just a few questions and remarks that may be useful to the authors as they think about some minor revisions and next steps:

2. Section 3.2, Lines 215 onward: I found it interesting that several of the carbon-terms dropped out due to their relatively negligible contributions. [CO2] effects on crops (and ecosystems!), and their nonlinear interactions with other changing climate parameters, are still highly uncertain. Crop models also display much variability in their respective [CO2] responses. I noticed that for the simulations emulating HadGEM responses [CO2] was held fixed or not varying with other parameters. Since the authors are emulating, and evaluating against, GGCM outputs, if the GGCMs do not display [CO2], then it follows that neither will the emulators I suppose. However, I wonder if the authors could further comment on this: the fact that [CO2] was negligible for the emulation does not necessarily mean the effects are negligible in reality, correct? I don’t see this discussed much elsewhere in the manuscript, so having a bit more commentary on this, with respect to [CO2] and/or more generally, would be useful as readers consider the terms of your model emulator.

We agree that it is highly interesting that the higher-order interaction terms could be dropped for most models, and hope that this will be the subject of a follow-up paper. This type of finding is part of what makes emulators powerful as a diagnostic tool of model behavior. Two models (PROMET and JULES) required the higher order CO2 interactions for accurate emulation, and it would be interesting to understand why.

Note that the magnitude of the pure CO2 terms is very large. The CO2 response is critical and results in large yield changes. (See Figure 10 for example).

In the HadGEM simulations shown in Figure 9, we held out CO2 precisely because the crop CO2 response is large and the purpose of this exercise was to examine the fidelity of the emulators' temperature / precipitation response. Figure 9 examines whether an emulator trained on the GGCMI Phase II database, which allows for no changes in climate variability, can accurately reproduce crop yields under actual climate model output, that may involve some changes in variability as well as means. We agree that this issue is under-discussed and have now added text to explain this more carefully. We now explicitly note that the CO2 response in LPJmL is so large that it almost completely negates the damages caused by higher temperatures, and that we hold it out to isolate the temperature-driven response.

Text added to section 4.3, line 396 - 415.

3. Section 4 and elsewhere: This comment is not just relegated to Section 4, but I'm more generally trying to parse out the relative contributions of climate variability and mean climate change, and the arguments provided in the paper that support emulation of the latter. I think there are two types of "variability" (admittedly not the best word, perhaps more "characteristics" other than the climatological mean yield) that the authors address that might be clarified just a bit more in the Discussion to avoid any confusion. Firstly, in Section 2.2, the authors make the case that year-to-year variability is structurally different than simulation of the climatological mean yields, and that the former doesn't preclude the latter, correct? The authors also highlight (I think) that the emulators are not suitable for a full interpretation of interannual variability and extremes, particularly highly non-linear interactions between climate (and other) parameters, despite the higher order terms of the emulation (which, as the authors note, are geared towards emulating climatological means).

Correct. Climatological mean yields are closely related to climatological mean temperature. Year-to-year yields are driven by weather factors other than (or in addition to) mean temperature. We show in Figure1 that regressing on growing-season mean temperature and climatological yield responses does not allow capturing the year-to-year variations. Presumably capturing both effects simultaneously in a single statistical model would require different regressors than growing-season mean temperature. We have added language to clarify this point and to clarify that the emulators should not be used in the study of responses to short-term extremes within the growing season.

Additional text pertaining to the choice of mean yields has been added to Section 2.2, Lines 118 - 125.

4. Secondly, in Section 4.3, the authors demonstrate that potential shifts in the distributions of climate parameters do not impact the climatological yield emulations and the results still compare well with the GGCMs (Figure 9). This fact – this shift in distribution and potential changes in variability that result from it (which is where the readers' mind might go, as mine certainly did) – is I believe distinct from the discussion of year-to-year variability discussed above.

This is a separate but related point. Because our emulators are trained on climate simulations with uniform offsets and no change in the other moments of the distribution, we felt the need to show that the emulation could still faithfully capture the response of crop models when driven by a climate model projection, which includes some changes in variability. Because our emulator is not trained on any aspect

of year-over-year variability, it was important to ask whether changes in variability in climate models might be so large and impactful for crops that they dominated the effects of mean changes and made the GGCMI emulators not useful. By showing that the emulated yield change is equal to the change simulated under a climate projection with the same mean temperature shift, we demonstrate that any variability changes in climate projections are not large/impactful enough to invalidate the GGCMI emulators. We have added language to our discussion to clarify this point.

Text modified in section 4.3, lines 396 - 415.

5. I appreciate that the authors have provided detailed explanations of their approach, treatment, and findings wrt to considering these variability and distributional changes. Still, there's a lot of material here to keep track of, and I think it may be useful to reiterate each of the above points clearly in the Discussion (particularly if I've mistakenly represented it, as I think this may be an example of reader confusion!). For example, there is a sentence in the Discussion that minimizes the impact of future variability (par- ticularly at the aggregate level – around Line 445), particularly in the area aggregate, and I think this is in reference to the findings in Section 4.3 However, this doesn't mean that interannual variability, or extremes or nonlinear interactions, won't be impactful tofuture (or current) crop impacts.

We agree that a recap would be very helpful; thank you for the suggestion. We have expanded the discussion as suggested.

Recap added to section 6, lines 496 - 527.

6. Discussion: Lastly, I think the major point of this paper is to provide these emulator frameworks as an alternative to climate-crop assessments with the full process- based GGCMs. I therefore understand the authors' approach to evaluate the emulators against the GGCMs – this is quite reasonable.

   It might be helpful, though, to take one step beyond this and compare to some observed historical yield changes. I would not expect this to be better than the GGCMs, and such evaluations have already been done for the GGCMs, so I would expect to see a similar response (and this is notwithstanding the applicability and veracity of comparison products). However, I don't think I've seen such an evaluation for GGCMI Phase 2 yet (I expect one is planned), and so pre-empting this with a comparison of the emulators may just be useful to have on hand. If this could be done and stuck into Supplementary, it would be a useful figure for the community moving forward, rather than having to show an intermediary figure of emulator-GGCM comparisons.

Unfortunately this type of validation is impossible with our current approach. On the decadal timescale, changes in management outweigh the effects of climate, and climate-driven mean yield changes in the historical record are impossible to disentangle from management changes. On the yearly timescale, as discussed above, the emulators are not appropriate for reproducing short-term variations, and so we also cannot use them to faithfully represent historical yearly yield anomalies (detrended from management changes).

The performance of the GGCMI Phase 2 crop models is addressed in the GMD companion paper (Franke et al. 2020a), using the standard evaluation approach based on the year-over-year time-series correlation

with FAO statistics (see Müller et al. 2017). However, this time-scale is not addressed by the emulators of the crop models, and so the emulators cannot be treated similarly.

## References


Elliott J, Müller C, Deryng D, Chryssanthacopoulos J, Boote KJ, Büchner M, Foster I, Glotter M, Heinke J, Iizumi T, Izaurralde RC, Mueller ND, Ray DK, Rosenzweig C, Ruane AC, and Sheffield J. 2015, The Global Gridded Crop Model Intercomparison: data and modeling protocols for Phase 1 (v1.0), Geoscientific Model Development, 8, 261-277, doi: 10.5194/gmd-8-261-2015.

Franke J, Müller C, Elliott J, Ruane AC, Jägermeyr J, Balkovic J, Ciais P, Dury M, Falloon P, Folberth C, Francois L, Hank T, Hoffmann M, Izaurralde RC, Jacquemin I, Jones C, Khabarov N, Koch M, Li M, Liu W, Olin S, Phillips M, Pugh TAM, Reddy A, Wang X, Williams K, Zabel F, and Moyer E. 2020, The GGCMI Phase II experiment: global gridded crop model simulations under uniform changes in CO2, temperature, water, and nitrogen levels (protocol version 1.0), Geoscientific Model Development, 13, 2315-2336, doi: 10.5194/gmd-13-2315-2020.

Iizumi T, Kim W, and Nishimori M. 2019, Modeling the Global Sowing and Harvesting Windows of Major Crops Around the Year 2000, Journal of Advances in Modeling Earth Systems, 11, 99-112, doi: 10.1029/2018MS001477.

Minoli S, Egli DB, Rolinski S, and Müller C. 2019, Modelling cropping periods of grain crops at the global scale, Global and Planetary Change, 174, 35-46, doi: 10.1016/j.gloplacha.2018.12.013.

Müller C, Elliott J, Chryssanthacopoulos J, Arneth A, Balkovic J, Ciais P, Deryng D, Folberth C, Glotter M, Hoek S, Iizumi T, Izaurralde RC, Jones C, Khabarov N, Lawrence P, Liu W, Olin S, Pugh TAM, Ray DK, Reddy A, Rosenzweig C, Ruane AC, Sakurai G, Schmid E, Skalsky R, Song CX, Wang X, de Wit A, and Yang H. 2017, Global gridded crop model evaluation: benchmarking, skills, deficiencies and implications, Geoscientific Model Development, 10, 1403-1422, doi: 10.5194/gmd-10-1403-2017.

Nelson GC, Valin H, Sands RD, Havlík P, Ahammad H, Deryng D, Elliott J, Fujimori S, Hasegawa T, Heyhoe E, Kyle P, Von Lampe M, Lotze-Campen H, Mason d'Croz D, van Meijl H, van der Mensbrugghe D, Müller C, Popp A, Robertson R, Robinson S, Schmid E, Schmitz C, Tabeau A, and Willenbockel D. 2014, Climate change effects on agriculture: Economic responses to biophysical shocks, Proceedings of the National Academy of Sciences, 111, 3274-3279, doi: 10.1073/pnas.1222465110.

Schewe J, Otto C, and Frieler K. 2017, The role of storage dynamics in annual wheat prices, Environmental Research Letters, 12, 054005, doi: 10.1088/1748-9326/aa678e.

Wiebe K, Lotze-Campen H, Sands R, Tabeau A, van der Mensbrugghe D, Biewald A, Bodirsky B, Islam S, Kavallari A, Mason-D'Croz D, Müller C, Popp A, Robertson R, Robinson S, van Meijl H, and Willenbockel D. 2015, Climate change impacts on agriculture in 2050 under a range of plausible socioeconomic and emissions scenarios, Environmental Research Letters, 10, 085010, doi: 10.1088/1748-9326/10/8/085010.

Ruane, Alex C, Cynthia Rosenzweig, Senthold Asseng, Kenneth J Boote, Joshua Elliott, Frank Ewert, James W Jones, et al. 2017, An AgMIP Framework for Improved Agricultural Representation in Integrated Assessment Models. Environmental Research Letters, 12: 125003, doi:10.1088/1748-9326/aa8da6.

Additional modifications not directly related to reviewer comments:

- Phase II changed to Phase 2 (throughout)
- carbon changed to carbon dioxide (throughout)
- soy changed to soybean (throughout)
- Some spelling errors fixed
- Substantial re-ordering of Section 6 to more clearly address the reviewer comments and provide a recap as suggest by reviewers
- Some light language editing for clarity throughout

# The GGCMI ~~phase II~~ Phase 2 emulators: global gridded crop model responses to changes in $CO_2$, temperature, water, and nitrogen (version 1.0)

James A. Franke[1,2], Christoph Müller[3], Joshua Elliott[2,4], Alex C. Ruane[5], Jonas Jägermeyr[4,2,3,5], Abigail Snyder[6], Marie Dury[7], Pete D. Falloon[8], Christian Folberth[9], Louis François[7], Tobias Hank[10], R. Cesar Izaurralde[11,12], Ingrid Jacquemin[7], Curtis Jones[11], Michelle Li[2,13], Wenfeng Liu[14,15], Stefan Olin[16], Meridel Phillips[5,17], Thomas A. M. Pugh[18,19], Ashwan Reddy[11], Karina Williams[8,20], Ziwei Wang[1,2], Florian Zabel[10], and Elisabeth J. Moyer[1,2]

[1]Department of the Geophysical Sciences, University of Chicago, Chicago, IL, USA
[2]Center for Robust Decision-making on Climate and Energy Policy (RDCEP), University of Chicago, Chicago, IL, USA
[3]Potsdam Institute for Climate Impact Research, Member of the Leibniz Association, Potsdam, Germany
[4]Department of Computer Science, University of Chicago, Chicago, IL, USA
[5]NASA Goddard Institute for Space Studies, New York, NY, United States
[6]Joint Global Change Research Institute, Pacific Northwest National Laboratory, College Park, MD, USA
[7]Unité de Modélisation du Climat et des Cycles Biogéochimiques, UR SPHERES, Institut d'Astrophysique et de Géophysique, University of Liège, Belgium
[8]Met Office Hadley Centre, Exeter, United Kingdom
[9]Ecosystem Services and Management Program, International Institute for Applied Systems Analysis, Laxenburg, Austria
[10]Department of Geography, Ludwig-Maximilians-Universität, Munich, Germany
[11]Department of Geographical Sciences, University of Maryland, College Park, MD, USA
[12]Texas Agrilife Research and Extension, Texas A&M University, Temple, TX, USA
[13]Department of Statistics, University of Chicago, Chicago, IL, USA
[14]EAWAG, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland
[15]Laboratoire des Sciences du Climat et de l'Environnement, LSCE/IPSL, CEA-CNRS-UVSQ, Université Paris-Saclay, F-91191 Gif-sur-Yvette, France.
[16]Department of Physical Geography and Ecosystem Science, Lund University, Lund, Sweden
[17]Earth Institute Center for Climate Systems Research, Columbia University, New York, NY, USA
[18]School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham, UK.
[19]Birmingham Institute of Forest Research, University of Birmingham, Birmingham, UK.
[20]Global Systems Institute, University of Exeter, Laver Building, North Park Road, Exeter, EX4 4QE, UK

**Correspondence:** James Franke (jfranke@uchicago.edu)

**Abstract.** Statistical emulation allows combining advantageous features of statistical and process-based crop models for understanding the effects of future climate changes on crop yields. We describe here the development of emulators for nine process-based crop models and five crops using output from the Global Gridded Model Intercomparison Project (GGCMI) Phase~~II~~ 2. The GGCMI Phase~~II~~ 2 experiment is designed with the explicit goal of producing a structured training dataset for emulator development that samples across four dimensions relevant to crop yields: atmospheric carbon dioxide ($CO_2$) concentrations, temperature, water supply, and nitrogen inputs (CTWN). Simulations are run under two different adaptation assumptions: that growing seasons shorten in warmer climates, and that cultivar choice allows growing seasons to remain

fixed. The dataset allows emulating the climatological mean yield response ~~without relying on interannual variations~~of all models with a simple polynomial in mean growing-season values. Climatological mean yields are a central metric in climate change impact analysis; we show ~~that these are quantitatively different. Climatological mean yield responses can be readily captured with a simple polynomial in nearly all locations, with errors significant only in some marginal lands where crops are not currently grown.~~ here that they can be captured without relying on interannual variations. In general, emulation errors are negligible relative to differences across crop models or even across climate model scenarios; errors become significant only in some marginal lands where crops are not currently grown. We demonstrate that the resulting GGCMI emulators can reproduce yields under realistic future climate simulations, even though the GGCMI Phase~~II~~ 2 dataset is constructed with uniform CTWN offsets, suggesting that the effects of changes in temperature and precipitation distributions are small relative to those of changing means. The resulting emulators therefore capture relevant crop model responses in a lightweight, computationally tractable form, providing a tool that can facilitate model comparison, diagnosis of interacting factors affecting yields, and integrated assessment of climate impacts.

## 1 Introduction

Improving our understanding of the impacts of future climate change on crop yields is critical for global food security in the twenty-first century. Projections of future yields under climate change are generally made with one of two approaches: either process-based models, which simulate the process of photosynthesis and the biology and phenology of individual crops, or statistical models, which use historical weather and yield data to capture relationships between observed crop yields and major drivers. Process-based crop models provide some advantages, including capturing the direct effects of $CO_2$ fertilization and allowing projections in areas where crops are not currently grown. However, they are computationally expensive, and can be difficult or impossible to directly integrate into integrated climate change impacts assessments. Statistical crop models can only capture crop responses under the range of current conditions, but have several advantages: they implicitly include management and behavioral practices that are difficult to model explicitly, and they are typically simple analytical expressions that are easily implemented by downstream impact modelers. Both types of models are routinely used, and comparative studies have concluded that when done carefully, both approaches can provide similar yield estimates (e.g. Lobell and Burke, 2010; Moore et al., 2017; Roberts et al., 2017; Zhao et al., 2017; Liu et al., 2016a).

Statistical emulation allows combining some of the advantageous features of both statistical and process-based models. The approach involves constructing a "surrogate model" of numerical simulations by using their output as training data for a statistical representation (e.g. O'Hagan, 2006; Conti et al., 2009). Emulation is particularly useful in cases where simulations are complex and output data volumes are large, and has been used in a variety of fields, including hydrology (e.g. Razavi et al., 2012), engineering (e.g. Storlie et al., 2009), environmental sciences (e.g. Ratto et al., 2012), and climate (e.g. Castruccio et al., 2014; Holden et al., 2014). For agricultural impacts studies, emulation of process-based models allows capturing key relationships between input variables in a lightweight, flexible form that is compatible with economic studies. The resultant

statistical model can produce yield projections under arbitrary emissions scenarios and is an important diagnostic tool for model comparison and model evaluation.

Interest is rising in applying statistical emulation to crop models, and multiple studies have developed crop model emulators in the past decade. Early studies proposing or describing potential crop yield emulators include Howden and Crimp (2005); Räisänen and Ruokolainen (2006); Lobell and Burke (2010), and Ferrise et al. (2011). Studies developing single-model emulators include Holzkämper et al. (2012) for the CropSyst model, Ruane et al. (2013) for the CERES wheat model, and Oyebamiji et al. (2015) for the LPJmL model. More recently, emulators have begun to be used in the context of multi-model intercomparison, with multiple authors (Blanc and Sultan, 2015; Blanc, 2017; Ostberg et al., 2018; Mistry et al., 2017) using them to analyze the five crop models of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP). ISIMIP offers a relatively large training set – control, historical, and several Representative Concentration Pathway (RCP) scenarios using output from up to five climate models (Warszawski et al., 2014; Frieler et al., 2017) – and choices of emulation strategy differ. Blanc and Sultan (2015) and Blanc (2017) use historical and RPC8.5 scenarios, combine multiple climate model projections for RCP8.5, and regress across soil regions. Ostberg et al. (2018) use global mean temperature change (and $CO_2$) as regressors, and then pattern-scales to emulate local yields. Mistry et al. (2017) compare emulated and observed historical yields, using local weather data and a historical crop simulation. The constraints of the ISIMIP experiment mean that all these efforts do share important common features. All emulate annual crop yields along an entire scenario or scenarios, and all future climate scenarios are non-stationary, with important covariates (temperature and precipitation for example) evolving simultaneously.

An alternative approach to emulation involves construction of a "parameter sweep" training set, a collection of multiple stationary scenarios that systematically cover a range of input parameter values. A parameter sweep offers several important advantages for emulation over an experiment in which climate evolves over time. First, it allows separating the effects of different variables that affect yields but that are highly correlated in realistic future scenarios like those used in ISIMIP (e.g. $CO_2$ and temperature). Second, it allows making a distinction between year-to-year yield variations and climatological changes, which may involve different responses to the particular climate regressors used (e.g. Ruane et al., 2016). For example, if year-to-year yield variations are driven predominantly by variations in the distribution of temperatures throughout the growing period, and long-term climate changes are driven predominantly by additive mean shifts, then regressing on the mean growing period temperature will produce different yield responses at annual vs. climatological timescales.

Systematic parameter sweeps have begun to be used in crop model evaluation and emulation, with early efforts in 2014 and 2015 (Ruane et al., 2014; Makowski et al., 2015; Pirttioja et al., 2015), and several recent studies in 2018 and 2019 (Fronzek et al., 2018; Ruiz-Ramos et al., 2018; Snyder et al., 2019). These three studies sample multiple perturbations to temperature and precipitation, and two of the three add $CO_2$ as well, for a total of 132, 99, and 220 different combinations, respectively. All take advantage of the structured training set to construct emulators ("response surfaces") of climatological mean yields, omitting year-to-year variations. All the 2018–2019 papers have some limitations, however, for assessing global agricultural impacts, including that none evaluate responses in every grid cell globally. Two involve many crop models but only one crop (wheat) (Fronzek et al., 2018; Ruiz-Ramos et al., 2018) and cover only 1–4 individual sites. Snyder et al. (2019) analyzes

five crops over ∼1000 sites with individual site-specific crop models, and extrapolates in space to estimate mean latitudinal responses.

In this paper we describe a set of globally-gridded crop model emulators developed from the new parameter-sweep dataset of the Global Gridded Crop Model Intercomparison (GGCMI) Phase~~II~~ 2 effort. GGCMI Phase~~II~~ 2, a part of the Agricultural Model Intercomparison and Improvement Project (AgMIP) (Rosenzweig et al., 2013, 2014), provides the first near-global-coverage systematic parameter sweep of multi-model crop simulations consisting of up to 756 combinations in $CO_2$, temperature, water supply, applied nitrogen, and two different assumptions on growing season adaptation ("A0": none and "A1": retaining growing season length) (CTWN-A, Franke et al., 2020; Minoli et al., 2019b). The experiment is designed to allow diagnosing the impacts on crop yields of both individual factors and their joint effects, and to allow construction of crop model emulators. In ~~the~~ Section 2 following, we describe the training dataset~~(Section 2) , the~~ , including the GGCMI Phase 2 experimental protocol and model participation (Section 2.1) and the models' differing year-over-year and climatological mean responses (Section 2.2). Section 3 describes the statistical model used for emulation~~(Section 3),~~ , Section 4 evaluates measures of emulator fidelity~~(Section 4), and~~ , and Section 5 shows examples of preliminary results~~(Section 5)~~.

## 2 Training dataset

### 2.1 The GGCMI Phase~~II~~ 2 dataset

The GGCMI Phase~~II~~ 2 simulations are described in detail in Franke et al. (2020), but we summarize briefly here. The experiment involves nine different globally gridded crop models, each simulating multiple crops (maize, rice, soybean, and spring and winter wheat) across a systematic parameter sweep of as many as 756 combinations, each driven by a historical climate timeseries with systematic perturbations to $CO_2$, temperature, water supply, and nitrogen application (CTWN). The simulation protocol involves 4 levels of atmospheric $CO_2$, 7 of temperature, 9 of water supply, and 3 of applied nitrogen, and simulations are repeated for two adaptation scenarios: "A0" simulations assume no adaptation in cultivar choice, so that growing seasons shorten in warmer climates, and "A1" simulations assume that adaptation in cultivar choice maintains fixed growing seasons. The complete protocol for each modeling group involves up to 43,524 years of global simulated output for each crop. Because the computational demand is high, modeling groups were allowed to submit at various specified levels of participation, with the lowest recommended level of participation consisting of 20% of the maximum possible simulations. The mean participation level is 65%, but three models (APSIM-UGOE, EPIC-IIASA, and ORCHIDEE-crop) contributed data below the recommended threshold ($< 5\%$ of the full protocol) and are excluded here since they could not be robustly emulated. Table 1 shows the participating models and the number of simulation scenarios that each provides, and Supplemental Figure S1 shows model sampling density. See Franke et al. (2020) for the parameter combinations included by each model. Table 2 shows the specified input values; we sample across all parameter combinations.

Each individual crop model simulation is run for 31 years over historic weather for the period of 1981-2010, with added uniform perturbations to any of the CTWN variables. Historical weather is taken for most models from the AgMERRA (Ruane et al., 2015) historical daily climate data product, but the PROMET model uses the ERA-Interim reanalysis (Dee et al., 2011)

**Table 1.** Crop models included in GGCMI Phase~~II~~ 2 emulators and the number of CTWN-A (Carbon dioxide, Temperature, Water, Nitrogen, Adaptation) simulations performed for each model. The maximum number is 756 for A0 (no adaptation) experiments, and 648 for A1 (maintaining growing season length) experiments, since T0 is not simulated under A1. "N-Dim." indicates whether the models are able to represent varying nitrogen levels. Each model provides the same set of CTWN simulations across all its modeled crops, but some models omit individual crops. Table adapted from Franke et al. (2020). For clarity, three simulation models ~~included in~~ that submitted data to the GGCMI Phase~~II~~ 2 experiment (Franke et al., 2020) are not shown here, ~~those that~~ as they provided a training set too small to be used in emulation.

| Model (Key Citations) | Maize | Soybean | Rice | Winter wheat | Spring wheat | N dim. | Sims per crop (A0 / A1) |
|---|---|---|---|---|---|---|---|
| **CARAIB**, Dury et al. (2011); Pirttioja et al. (2015) | X | X | X | X | X | – | **252 / 216** |
| **EPIC-TAMU**, Izaurralde et al. (2006) | X | X | X | X | X | X | **756 / 648** |
| **JULES**, Osborne et al. (2015); Williams and Falloon (2015); Williams et al. (2017) | X | X | X | – | X | – | **252 / 0** |
| **GEPIC**, Liu et al. (2007); Folberth et al. (2012) | X | X | X | X | X | X | 430 / 181 |
| **LPJ-GUESS**, Lindeskog et al. (2013); Olin et al. (2015) | X | – | – | X | X | X | **756 / 648** |
| **LPJmL**, von Bloh et al. (2018) | X | X | X | X | X | X | **756 / 648** |
| **pDSSAT**, Elliott et al. (2014); Jones et al. (2003) | X | X | X | X | X | X | **756 / 648** |
| **PEPIC**, Liu et al. (2016b, c) | X | X | X | X | X | X | 149 / 121 |
| **PROMET**, Hank et al. (2015); Mauser et al. (2015); Zabel et al. (2019) | X | X | X | X | X | – | 261 / 232 |

and the JULES model uses a bias-corrected version of ERA-Interim, WFDEI (WATCH-Forcing-Data-ERA-Interim, Weedon et al., 2014) as these groups have specific sub-daily input data requirements. Temperature perturbations are applied as additive mean shifts, water supply as fractional multipliers to precipitation (except in the irrigated $W_{\infty}$ case), and $CO_2$ and nitrogen application levels are specified as fixed values. Models provide near-global output at 0.5 degree latitude and longitude resolution for each simulation year, including areas not currently cultivated. Crop models included here are not formally calibrated, given that there is no adequate calibration target for gridded global-scale crop model simulations. This may be a shortcoming if targeting absolute yield levels, but when focusing on relative yield changes, calibration can also have negative effects on model

**Table 2.** GGCMI Phase~~II~~ 2 input levels for the parameter sweep. Values for temperature and water supply are perturbations from the historical climatology. For water supply, perturbations are fractional changes to historical precipitation, except in the irrigated ($W_\infty$) simulations, which are all performed with the maximum beneficial levels of water. Bold font indicates the 'baseline' historical level. The full protocol samples across all parameter combinations for a total of 756 cases. Table repeated from Franke et al. (2020).

| Input variable | Tested range | Unit |
|---|---|---|
| [$CO_2$] (C) | **360**, 510, 660, 810 | ppm |
| Temperature (T) | -1, **0**, 1, 2, 3, 4, 6 | °C |
| Precipitation (W) | -50, -30, -20, -10, **0**, 10, 20, 30, (and $W_\infty$) | % |
| Applied nitrogen (N) | 10, 60, **200** | kg $ha^{-1}$ |
| Adaptation (A) | **A0: none**, A1: new cultivar to maintain original growing season length | - |

skill (Müller et al., 2017). In analyses where we distinguish yields over currently cultivated land, we use the harvested area masks of Portmann et al. (2010). (See Supplemental Figure S2 for maps of cultivated area.)

### 2.2 Climatological vs. year-to-year ~~response~~responses

~~We emulate the climatological mean response, because that is the response of interest~~ The central metric in assessments of climate change impacts ~~. The~~ on crop yields is the change in multi-annual means (e.g. Schlenker and Roberts, 2009; Challinor et al., 2014; Ros . Agricultural impacts assessments work with multi-annual yields, as their analysis frameworks require information on long-term effects (e.g. Nelson et al., 2014b; Stevanović et al., 2016; Wiebe et al., 2015; Hasegawa et al., 2018; Snyder et al., 2019). Changes in extremes or year-to-year ~~response can be significantly different from the forced climatological one, so we do not use information from year-to-year~~ variability ~~but instead emulate the~~ are other metrics of potential interest, but are often not explicitly considered in integrated climate change impact assessments or land-use change projections. For this reason we emulate the climatological mean response, i.e. the change in aggregated mean yield in each 30-year simulation. Emulation then becomes relatively straightforward, since changes in time-averaged yields are ~~also~~ considerably smoother than those in year-to-year yield response.

In the GGCMI Phase~~II~~ 2 simulation output dataset, year-to-year responses to weather are also often quantitatively distinct from responses to climatological shifts, with the discrepancy especially strong in wheat and rice. The difference in behavior is illustrated in Figure 1, which shows irrigated and rainfed maize and wheat in representative locations~~; open circles and black lines show the climatological mean response, and solid circles and colored lines the responses for the 30 individual years in individual scenarios~~. When discrepancies are large, year-to-year responses are generally stronger than climatological ones, but exact responses differ by crop and region and even by model within GGCMI Phase~~II.~~ 2.

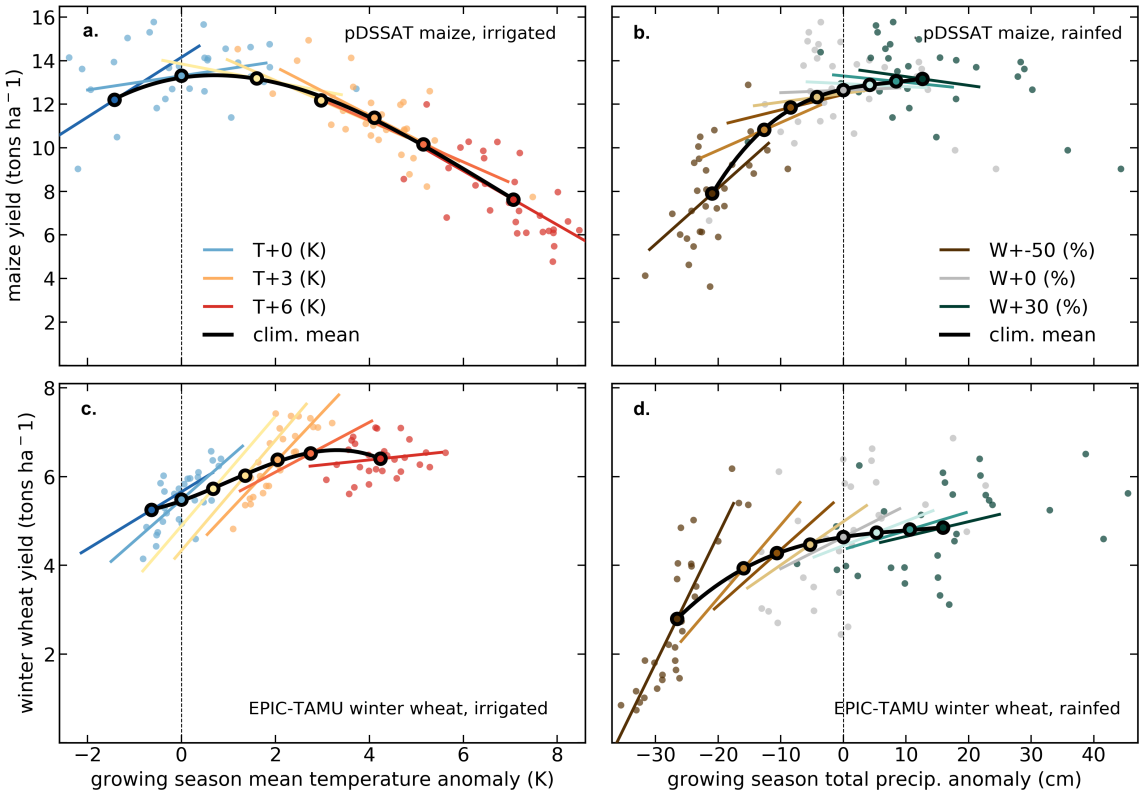


**Figure 1.** Example showing distinction between crop yield responses to year-to-year and climatological mean shifts in climate variables, showing representative high-yield regions for maize in pDSSAT (northern Iowa, top row) and winter wheat in EPIC-TAMU (France, bottom row). Left column (**a & c**) shows irrigated crops, all temperature cases with other variables held at baseline values, and right column (**b & d**) shows rainfed crops, all precipitation cases. Figure shows A0 output, in which growing seasons shift under future climate, so local growing-season temperature changes can differ from prescribed uniform offsets: for example, a 6 K applied uniform warming results in a growing season temperature warmer by ∼7 K for maize in Iowa (top right), but by less than 6 K for wheat in France (bottom right). Open black circles mark climatological mean yields and bold black lines show a 3rd order polynomial fit through them. Colored lines show linear regressions (by orthogonal distance regression) through the 30 annual yields of each parameter case. Colored circles show annual yields for selected cases. Differences in slopes of colored and black lines mean that responses to year-to-year fluctuations differ from those to longer-term climate shifts. Differences are generally stronger for wheat (bottom) than maize (top). Note that for rain-fed crops, slope differences in this representation could also result from correlated precipitation and temperature fluctuations in the baseline timeseries, but P-T correlations do not contribute to the effects shown here. Such correlations would complicate emulations based on year-to-year yields but would not necessarily bias them.

While differences in responses at different timescales can arise for many reasons, including memory in the crop model or lurking covariates, the most likely explanation here is that the regressors used, mean growing-season temperature or precip-
135 itation, do not fully describe the conditions that affect crop yields. The mean growing-season value is only a proxy for the

distribution of daily climatic conditions that crops are sensitive to, and present-day variations between years can be very different from future forced changes. That is, present-day variations in growing-season *means* from year to year may be associated with changes in growing-season *distributions* that are unrelated to ~~any~~ changes in future warmer climates: that is, a warm year at present may be quite different from a warm year in the future (e.g. Ruane et al., 2016). Changes in temperature distributions have been shown to strongly affect crop yields (e.g. Hansen and Jones, 2000; Gadgil et al., 2002), though precipitation effects should be smaller since crops respond not to rainfall but to soil moisture, which integrates over weeks or even months (e.g. Potter et al., 2005; Glotter et al., 2014; Challinor et al., 2004).

A second factor of importance is that any nonlinearity in crop responses will itself lead to a distinction between climatological and year-to-year fits, even if distributional differences are negligible. Given the interannual variations in the climate timeseries, the mean annual yield response to a perturbation is not the same as the response of the climatological mean yield. The effect of nonlinearity may be particularly relevant for precipitation, since model crop yields drop steeply and nonlinearly with increasing dryness. (Crop yields should drop under excess precipitation as well, but process-based models do not capture losses in saturated conditions well (Glotter et al., 2015; Li et al., 2019).)

In the GGCMI Phase~~II~~ 2 experiment, the imposed perturbations involve no changes in underlying distributions. The choice is reasonable, since climate models do not agree on distributional changes. Most models do project small mean increases in growing-season temperature variability in cultivated areas, and can produce substantial local changes, but models disagree on spatial patterns. For example, in models of the Coupled Model Intercomparison Project Phase 5 (~~CMIP-5~~CMIP5) archive, in the the high-end RCP (Representative Concentration Pathway) 8.5 climate projections to the year 2100 (Riahi et al., 2011), growing season daily maximum temperature variability over currently cultivated rice areas (weighted by production) increases by 10% in HadGEM2-ES but only by 0.4% in MIROC-ESM-CHEM. (See Supplemental Section S2.) We therefore explicitly test the assumption that distributional changes are not consequential for climatological mean yields: in Section 4.3, we confirm that an emulator trained on the GGCMI Phase~~II~~ 2 dataset can successfully reproduce yield changes under a full climate model projection.

Note that even though distributions of climate variables are unchanged in the GGCMI Phase~~II~~ 2 simulations, the spread in annual yields still becomes wider in highly impacted climate states, because of the nonlinearity of yield responses (Figure 2). In the GGCMI Phase~~II~~ 2 dataset, all crops except rice show greater year-to-year yield variance in conditions of extreme climate stress. (Rice is typically irrigated and experiences no water stress in simulations.) Increased variance has been noted in previous studies. For example, Urban et al. (2012) used statistical models trained on present-day yields to find a projected future increase in yield variance of U.S. maize of 20% per degree K temperature rise. While the authors do not diagnose a specific cause of that increase, they discuss multiple potential mechanisms, including nonlinearity in responses.

## 3 Emulation

Emulation involves fitting individual regression models from GGCMI Phase~~II~~ 2 output for each crop and model and 0.5 degree geographic pixel; the regressors are the applied perturbations in $CO_2$, temperature, water, and nitrogen (CTWN). We discuss

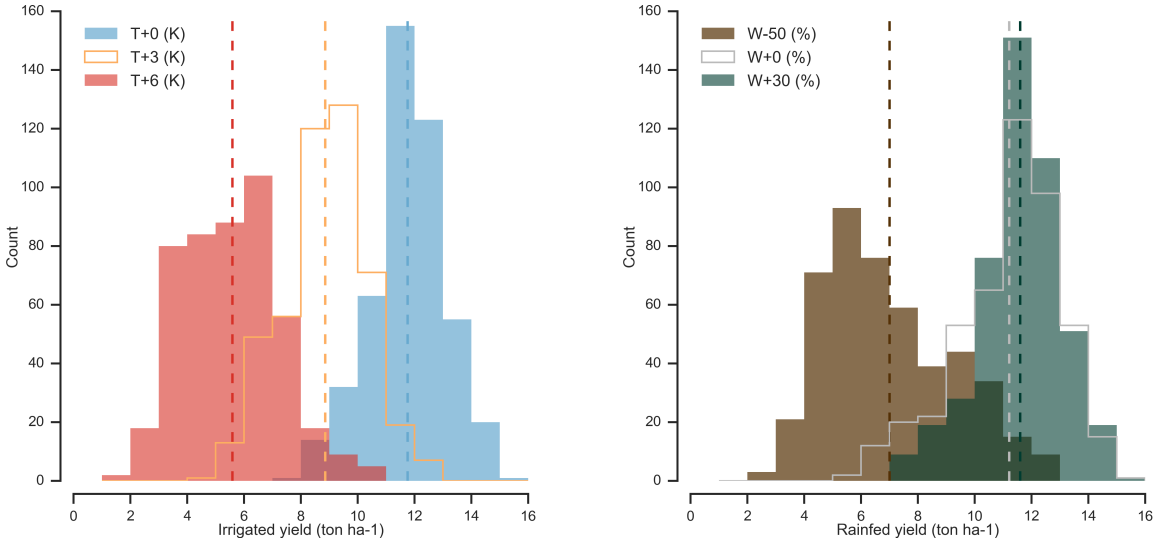


**Figure 2.** Example showing results of increased crop yield sensitivity to year-to-year climate variations under climate stress. Yield distributions are from examples of Figure 1, top row, of maize in Iowa, (**left**) for irrigated maize in scenarios of altered temperature and (**right**) for rainfed maize in scenarios of altered precipitation. Because yield sensitivities rise under strong warming or drying, distributions of year-to-year crop yields widen in T+6 and P-50% scenarios relative to present-day simulations, even though all input climate timeseries have identical variance for temperature. Note: precipitation changes have different variance since the perturbations are fractional.

here largely emulations of climatological mean crop yields with no growing season adaptation (A0 scenarios), but note that any output of the crop models can potentially be emulated. We provide separate emulations of irrigated and rainfed yields and applied irrigation water (pirrww in mm $yr^{-1}$) in both A0 and A1 scenarios, meaning that each model and crop combination results in six sets of regressions. See Supplemental Material Sections 3, 4, and 6 for these additional emulation cases.

### 3.1 Statistical model

For the statistical model of crop yields as a function of CTWN, we choose a relatively simple parametric model with a 3rd-order polynomial basis function (Equation 1). If the climatological mean response is relatively smooth, then a simpler form provides a reasonable fit that allows for some interpretation of resultant parameter weights. A ~~relativity~~ relatively simple parametric form also allows fast model emulation at the grid cell level, rather than requiring spatial aggregation. Emulating at the grid cell level preserves the spatial resolution of the parent models, and means that emulators indirectly includes any yield response to geographically distributed factors such as soil type, insolation, and the baseline climate.

The 3rd-order polynomial CTWN model of Equation 1 contains 34 terms~~(Equation 1)~~, since the $N^3$ term is omitted, as it cannot be fitted in a training set sampling only three nitrogen levels. To facilitate comparing emulators parameter by parameter, we hold this functional form across locations, crops, and models, other than several necessary distinctions: regressions for irrigated crops do not contain W terms, and regressions for models that do not sample the nitrogen levels omit the N terms.

Results shown throughout the paper use this full specification, but we also ~~investigate~~ show (in Section 3.2 below) ~~whether some~~ that for all but two models, 11 terms can be dropped without significant reduction in emulator fidelity. The higher specification of the 34 term model aids primarily in regions where crops are not currently grown. Most modeling groups submitted a sufficiently large training set that the 34-term model can be fit with standard ordinary least squares (OLS), but for models with lower sampling, it must be fit with a Bayesian Ridge regression method. (See Section 4 for evaluation of the fidelity of emulators constructed with Equation 1.)

$$
\begin{aligned}
Y \; = & \; K_1 \\
& + K_2 C + K_3 T + K_4 W + K_5 N + K_6 C^2 \\
& + K_7 CT + K_8 CW + K_9 CN + K_{10} T^2 + K_{11} TW \\
& + K_{12} TN + K_{13} W^2 + K_{14} WN + K_{15} N^2 \\
& + K_{16} C^3 + K_{17} C^2 T + K_{18} C^2 W + K_{19} C^2 N \\
& + K_{20} CT^2 + K_{21} CTW + K_{22} CTN + K_{23} CW^2 \\
& + K_{24} CWN + K_{25} CN^2 + K_{26} T^3 + K_{27} T^2 W \\
& + K_{28} T^2 N + K_{29} TW^2 + K_{30} TWN + K_{31} TN^2 \\
& + K_{32} W^3 + K_{33} W^2 N + K_{34} WN^2 + K_* N^3
\end{aligned}
\tag{1}
$$

We do not focus in this study on comparing other functional forms or non-parametric models. In general, ~~both~~ higher-order and interaction terms are expected to be important for representing crop yields. Higher order terms are needed because crop yield responses to weather are well-documented to be nonlinear: e.g. Schlenker and Roberts (2009) for T perturbations and He et al. (2016) for W (precipitation). Interaction terms are needed since the yield response is expected to depend on interactions between the major inputs. For example, Lobell and Field (2007) and Tebaldi and Lobell (2008) showed that in real-world yields (with C and N fixed), the joint distribution in T and W is needed to explain observed yield variance. Other observation-based studies have shown the importance of the interaction between W and N (e.g. Aulakh and Malhi, 2005), and between N and C (Osaki et al., 1992; Nakamura et al., 1997).

~~We do not focus in this study on comparing other functional forms or non-parametric models.~~ Some prior studies have used ~~other~~ even more complex statistical specifications in crop model emulation: for example, Blanc and Sultan (2015) and Blanc (2017) use a 39 term fractional polynomial and "borrow information across space" by fitting grid points simultaneously across soil region in a panel regression. The GGCMI Phase~~II~~ 2 dataset allows fitting our simple 3rd order polynomial form independently at each grid cell while still providing a satisfactory emulation for all models and crops. ~~(See Section 4 for evaluation of the fidelity of emulators constructed with Equation 1.)~~

### 3.2 Feature importance and reduced statistical model

Because a simpler statistical model may improve the interpretability of its parameter weights, we also develop a reduced 23-term version that is satisfactory for most models and crops (Equation 2, with the 11 removed terms shown in gray). To

identify terms that can be omitted, we apply a feature selection cross-validation process in which terms in the polynomial are tested for importance. Higher-order and interaction terms are successively added to the regression model, and in each case we calculate an aggregate mean absolute error (weighted by by currently cultivated area) and eliminate those terms that do not contribute significantly to reducing error. The procedure is illustrated in Figure 3. We develop our reduced statistical model by considering yields over currently cultivated land in three models: two that provided the complete set of 672 rainfed simulations, i.e. without the $W_{\infty}$ simulations, (pDSSAT, EPIC-TAMU), and one that provided the smallest training set (121 input combinations, PEPIC). Although models exhibit different absolute levels of error, all three agree remarkably well on feature importance, i.e. on which terms reduce error and which provide no predictive benefit. ~~(Agreement means that line slopes match~~ Agreement is indicated by matching line slopes in Figure 3. ~~)~~

~~Results of the feature selection process suggest that 11 terms can be omitted with negligible impact on emulator fidelity, producing the 23-term statistical model of Equation 2.~~

$$
\begin{aligned}
Y \; &= K_1 \\
&+ K_2 C + K_3 T + K_4 W + K_5 N + K_6 C^2 \\
&+ K_a CT + K_7 CW + K_8 CN + K_9 T^2 + K_{10} TW \\
&+ K_{11} TN + K_{12} W^2 + K_{13} WN + K_{14} N^2 \\
&+ K_* C^3 + K_* C^2 T + K_* C^2 W + K_* C^2 N \\
&+ K_* CT^2 + K_* CTW + K_* CTN + K_* CW^2 \\
&+ K_* CWN + K_{15} CN^2 + K_{16} T^3 + K_{17} T^2 W \\
&+ K_* T^2 N + K_{18} TW^2 + K_{19} TWN + K_{20} TN^2 \\
&+ K_{21} W^3 + K_{22} W^2 N + K_{23} WN^2 + K_* N^3
\end{aligned}
\tag{2}
$$

The eliminated terms include many of those in C: the cubic; the CT, CTN, CTW, and CWN interaction terms; and all higher order interaction terms in C. Finally, we eliminate one 2nd-order interaction term in W and two in T. Implications of this choice include that nitrogen interactions are complex and important, and that water interaction effects are more nonlinear than those in temperature. Note that some terms that did not reduce the aggregate error must still be included if a higher order version of that term provides benefit: for example, including the $T^3$ term requires also retaining $T^2$ and $T$ terms. The reduced-form emulator is acceptable across currently cultivated land for all model and crop combinations other than JULES ~~soy~~ soybean and spring wheat and PROMET ~~soy~~ soybean and rice. These cases involve yield responses that benefit strongly from inclusion of higher order carbon dioxide interaction terms. Additional terms in the statistical model also help emulation in some geographic locations outside of currently cultivated regions, where yield responses are often non-standard. ~~(~~See Supplemental Material Section 7 for evaluation of the fidelity of emulators constructed with Equation 2 and for more details on JULES and PROMET. ~~)~~

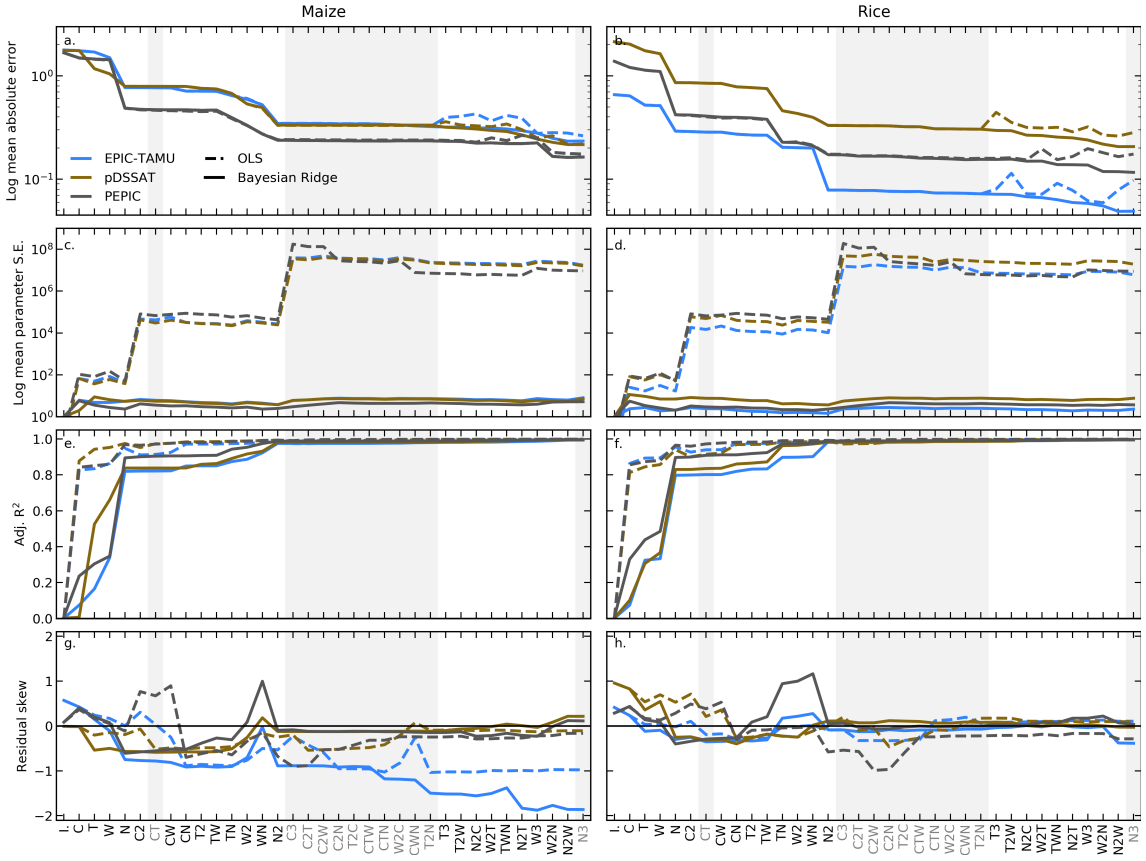


**Figure 3.** Illustration of results from the polynomial feature selection process for three different crop models (colors), for all grid cells with more than 1000 ha cultivated for maize (**left**) and rice (**right**). Solid lines are Bayesian Ridge regression results and dashed lines those for standard OLS. Rows show four metrics of fit quality and x axes the terms successively tested in the statistical model, sequentially added to the model in order from left to right. Terms that do not reduce the aggregate error are marked in gray and are not included in the final model. **a & b:** log mean absolute error between emulated yield and simulated values calculated with a three fold cross validation process, where the emulator is trained on two thirds of the data and predicts the remaining third. **c & d:** log mean standard parameter error. The Bayesian Ridge method strongly reduces parameter error and results in more stable estimates. **e & f:** adjusted $R^2$ score for the fit at each model specification. **g & h:** distribution of the residuals. Skewness is low at the high model specifications tested in all model cases other than EPIC-TAMU maize.

### 3.3 Model fitting

To fit the parameters $K$, we use a Bayesian Ridge regularization method (MacKay, 1991) rather than ~~standard~~ ordinary least squares (OLS). The Bayesian Ridge method reduces volatility in parameter estimates when the sampling is sparse, by weighting parameter estimates towards zero, allowing the use of a consistent functional form across all models and locations. The choice slightly reduces mean absolute error for some of the high-order interaction terms in the model (Figure 3, top row) but drastically reduces standard parameter error in the model by stabilizing the estimates (Figure 3, third row). The estimation method scores

relatively lower on adjusted $R^2$ for the simplest parameter specifications, but quickly reaches parity with the OLS. We use adjusted $R^2$ as a metric because additional terms are penalized (Equation 3, where $n$ is the number of samples and $k$ is the number of features):

$$R^2_{adj} = 1 - \frac{(n-1)\cdot(1-R^2)}{n-k} \tag{3}$$

We use the implementation of the Bayesian Ridge estimator from the scikit-learn package in Python (Pedregosa et al., 2011).

An additional diagnostic of fit quality is the distribution of residuals: normally or near-normally distributed residuals imply that errors around the fit are random and unbiased. When fitting Equation 1 to the GGCMI Phase~~II~~ 2 dataset, the distribution of the residuals depends on the number of features included in the regression, the method for estimating the parameters, and the target distribution in the training set. The residuals are only normally distributed (pvalue > 0.05 in the Shapiro–Wilk test) for a single model, PEPIC, for any specification tested here, but their skew is relatively small except in a single case, EPIC-TAMU maize (Figure 3, fourth row). While including higher-order terms in the statistical model generally reduces residual skew, for EPIC-TAMU maize it increases skew instead, but also reduces the error in cross-validation, which we consider more important in the context of emulation. The residual distribution suggests that projections using the EPIC-TAMU maize emulator will tend to be biased high, but in practice the overall magnitude of these errors is below 2% of yield changes. (See Section 4.2.)

## 4 Emulator evaluation

In this section we show illustrations of GGCMI model yield responses to climate perturbations and evaluate the ability of our emulators to reproduce them. Model emulation with the parametric method used here requires that crop yield responses be sufficiently smooth and continuous to allow fitting with a relatively simple functional form; in Section 4.1 we show that this condition largely holds in the GGCMI Phase~~II~~ 2 simulations. In section 4.2 we evaluate metrics of emulator performance and show that emulation errors – discrepancies between emulation and simulation – are generally small, especially when compared to the differences across crop models or to projected yield changes. ~~Emulation errors become problematic only in certain,~~ We use the term *error* because, under the "perfect" model emulation approach, we take the simulation output to be perfect ground truth. We evaluate two separate error metrics, one more loose that incorporates information about the inter-model uncertainty, and one more stringent that tests out of sample prediction error within an individual model. For both metrics, emulation error is generally small other than in limited geographic locations, usually where crops are not currently grown. ~~We analyze here results using the 34-term polynomial of Equation 1; see Supplemental Material Section 7 for analogous analysis of the 23-term polynomial of Equation 2.~~ Finally, in Section 4.3, we assess the emulator's ability to reproduce crop yields in a more realistic future simulation driven by a climate model projection, and find that ~~any effects of changes in climate variability not included in the GGCMI Phase II training set are generally small relative to the effects of mean changes~~its performance remains satisfactory. We analyze here results using the 34-term polynomial of Equation 1; see Supplemental Material Section 7 for analogous analysis of the 23-term polynomial of Equation 2.

### 4.1 Yield response

Crop yields show strong spatial differentiation across geographic regions, and emulators are able to readily reproduce these patterns. Figure 4 ~~illustrates the spatial yield pattern~~ shows one example of simulated and emulated yields under current climate~~for one crop and model (~~, using maize in LPJmL~~)~~. Absolute emulation errors for this model-crop combination are low – 99.8% of grid cells have errors below 0.5 tons ha$^{-1}$ – but emulation errors as a percentage of baseline yield can be large in areas with low potential yield and no current cultivation in the real world (e.g. the Sahara, Patagonia). These regions are not currently viable for agriculture and may never become viable even under extreme climate change. Emulator spatial skill varies across models and crops, with maize being the quantitatively easiest to emulate across all models and locations.

Yield responses to the four main drivers considered here (C, T, W, and N) are also quite diverse across locations, crops, and models, but in nearly all cases the local climatological mean responses are smooth enough to permit emulation with the functional form used here. Figure 5 illustrates the geographic diversity of responses within a single crop and model, for rainfed maize in pDSSAT. While the $CO_2$ responses (in ton ha$^{-1}$ ~~/ppm~~ppm$^{-1}$) are quite similar, the precipitation response is stronger in more arid locations and the nitrogen responses appear strongly location-dependent. The heterogeneity in response supports the choice of emulating at the grid cell level. In regions with current cultivation, yields evolve smoothly across the space sampled, and the polynomial fit captures the climatological-mean response to perturbations well. Emulators do perform poorly in a few regions that involve discontinuous or irregular yield responses. Poor performance is illustrated here with PROMET maize in northern Canada, which is too cold for maize at present in PROMET (0 ton ha$^{-1}$ yield), but which shows an abrupt rise to moderate yields once temperature rises by 4 degrees. Under these conditions, the 3rd order polynomial cannot fit the response, and errors are high. See Section 4.2 for additional discussion.

Crop yield responses in all models generally follow similar functional forms at any given location, though with a spread in magnitude (Figure 6, which shows rainfed maize in northern Iowa in a selection of GGCMI models). Absolute yield differences between models can be substantial because some models are uncalibrated. In general, models are most similar in their responses to temperature perturbations, and least similar to changes in $CO_2$. That is, $CO_2$ fertilization effects *within* a single model are consistent across locations, but $CO_2$ effects differ strongly *across* models.

Note that while the nitrogen dimension is important, it is also the most troublesome to emulate in the GGCMI Phase~~II~~ 2 experiment because of its limited sampling. The GGCMI Phase~~II~~ 2 protocol specified only three nitrogen levels (10, 60 and 200 kg N y$^{-1}$ ha$^{-1}$), so a third-order fit would be over-determined but a second-order fit can result in potentially ~~unphysical~~ non-physical results. Steep and nonlinear declines in yield with lower nitrogen levels mean that some regressions imply a peak in yield between the 100 and 200 kg N y$^{-1}$ ha$^{-1}$ levels (Figure 6, right). While reduced yields under high nitrogen levels are physically possible and could reflect over-application at particular times in the growing period, they are implausible at the magnitude shown here and likely an artifact of the fit. The Bayesian Ridge estimator mitigates the 'peak-decline effect' in the nitrogen dimension relative to ordinary least squares, but does not entirely remove it. The polynomial fit also cannot capture the well-documented saturation effect of nitrogen application (e.g. Ingestad, 1977) as accurately as would be possible with a non-parametric model.

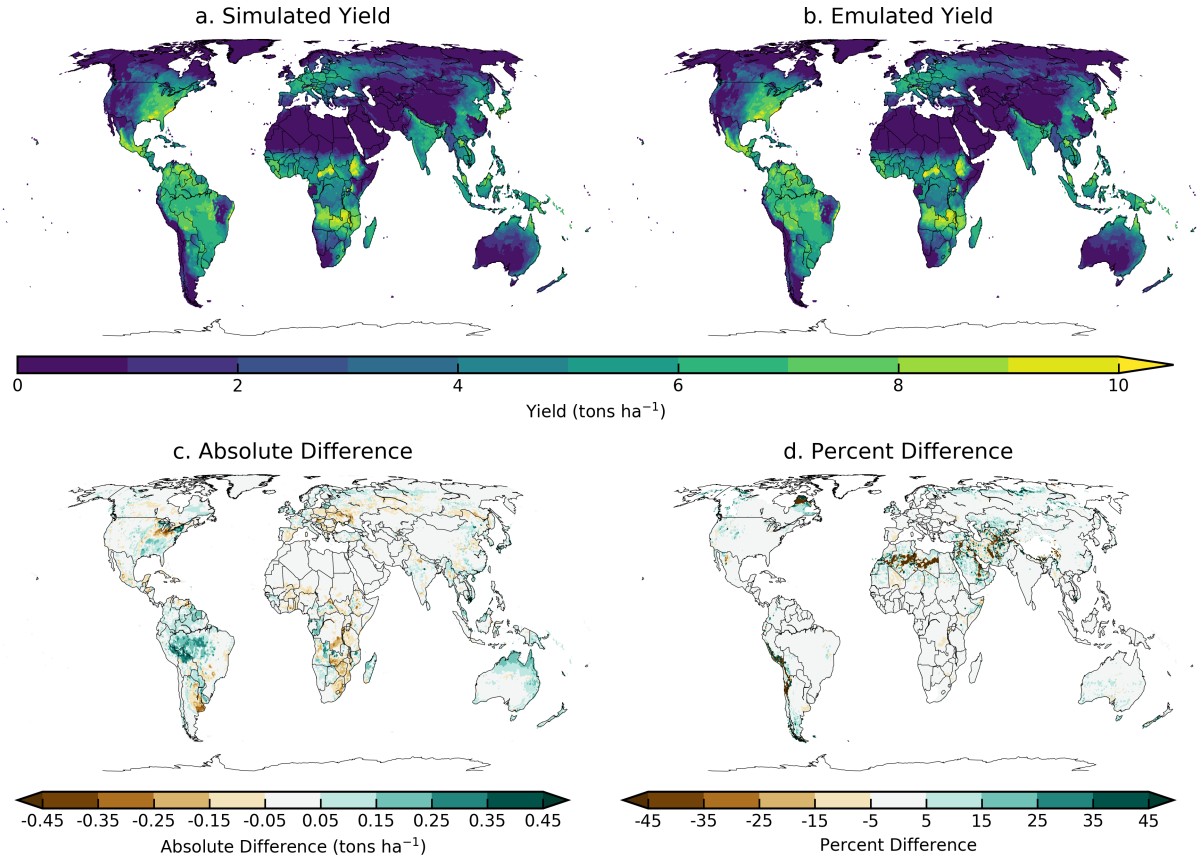


**Figure 4.** Illustration of spatial pattern in baseline yield successfully captured by the emulator: simulated (**a.**) and emulated (**b.**) yield under historical (1981-2010) conditions for rainfed maize from the LPJmL model. Absolute yield differences (**c.**) are less than 0.5 ton ha$^{-1}$ in almost all (99.8%) grid cells across the globe. Percent difference (from simulated baseline, **d.**) is below 5% in most (75%) grid cells currently cultivated in the real world. Approximately 7% of all grid cells, but only 3% of currently cultivated grid cell, have emulated yields that differ from the baseline simulation by more than 20%. Notable exceptions include areas with very low simulated baseline yield, including for example the Sahara, the Andes, and northern Quebec. Percent error weighted by cultivation area globally is essentially zero (see also Table 3). Performance varies by crop and model. See Supplemental Figures in Section 8 for more examples.

### 4.2 Emulator performance metrics

Our emulators collectively consist of nearly 3 million individual regressions, so developing concise performance metrics poses a challenge. No general agreed-upon criteria exist for defining an acceptable crop model emulator, so we present two different metrics below, one relatively loose and one more stringent. Both metrics assess the ability of the emulator to reproduce simulated crop yields in the GGCMI Phase~~II~~ 2 experiment. In this section we show only results from emulators based on the 34-term Equation 1; see Supplemental Material Section 7 for analogous assessment of emulators based on the 23-term Equation 2.

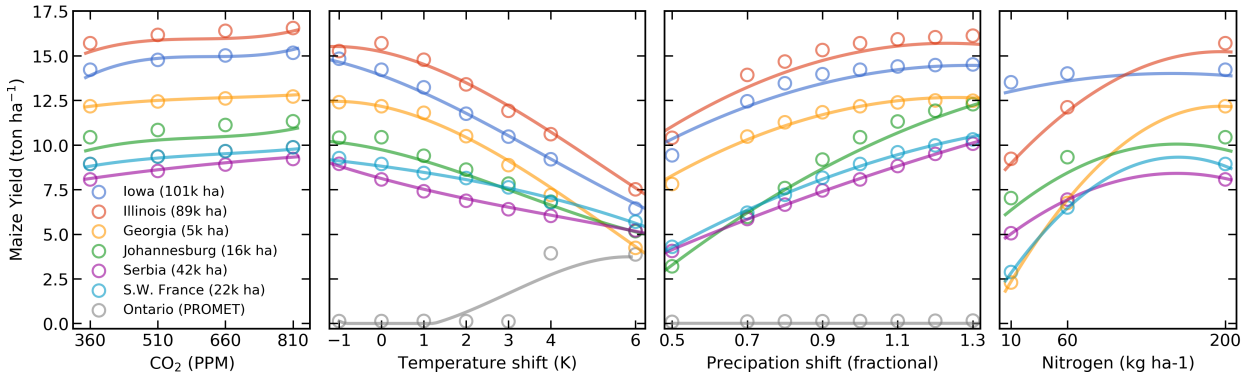


**Figure 5.** Illustration of spatial variations in yield response, which are successfully captured by the emulator. Panels show simulations (points) and emulations (lines) of rainfed maize in the pDSSAT model in six example locations selected to represent high-cultivation areas around the globe. Legend includes hectares cultivated in each selected grid cell. Each panel shows variation along a single variable, with others held at baseline values. Dots show climatological mean yields and lines the results of the full 4D emulator of Equation 1. In general the climatological response surface is sufficiently smooth that it can be represented within the sampled variable space by the simple polynomial used in this work. In some cases extrapolation would produce misleading results, and the emulator fails in conditions where yield response changes abruptly. Failure is illustrated here by rainfed maize in north-central Ontario for the PROMET model (in gray), which shows present-day yields of zero rising abruptly if temperature warms by 4 degrees.

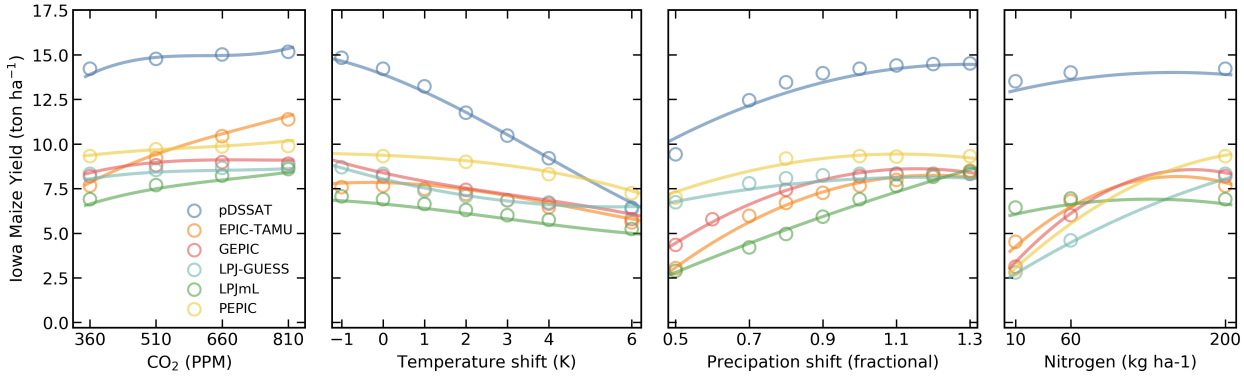


**Figure 6.** Illustration of variations in yield response across models, again successfully captured by the emulator. Panels show simulations and emulations from six representative GGCMI models for rainfed maize in the same Iowa grid cell shown in Figure 5, with the same plot conventions. Three models (PROMET, JULES, and CARAIB) that do not simulate the nitrogen dimension are omitted for clarity. Models are uncalibrated, producing spread in absolute yields. While most model responses can readily emulated with a simple polynomial, some response surfaces diverge slightly from the polynomial form, producing emulation error (e.g. pDSSAT here, for water), but resulting error generally remains small relative to differences across models.

325 *1. Normalized error.* We take as our first metric what we term the "normalized error", which compares the fidelity of an emulator to the inter-model spread. For a multi-model comparison exercise like GGCMI Phase~~I~~ 2, a reasonable though loose

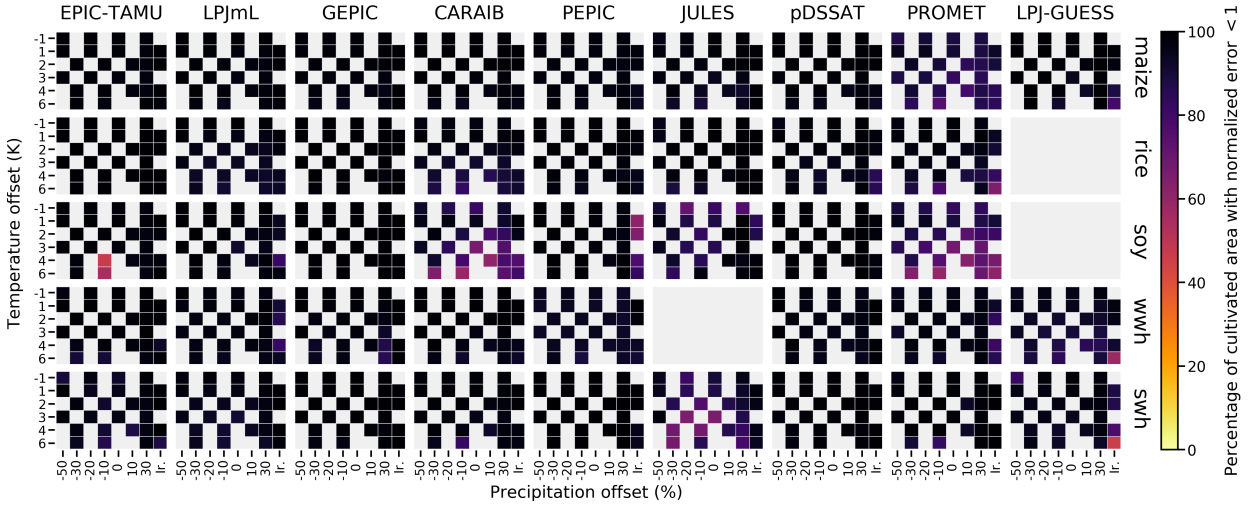


**Figure 7.** Assessment of emulator performance over currently cultivated areas based on normalized error (Equations 5). We show performance of all 9 models emulated, over all crops and all sampled T and W inputs ("ir." indicates the irrigated $W_\infty$ setting), but with $CO_2$ and nitrogen held fixed at baseline values. Large columns are crops and large rows models; squares within are T, W scenario pairs. Colors denote the fraction of currently cultivated hectares ("area frac") for each crop with normalized area $e$ less than 1 indicating the ~~the~~ error between the emulation and simulation less than one standard deviation of the ensemble simulation spread. Of the possible 63 scenarios at a single $CO_2$ and N value, we consider only those for which all 9 (8 for rice, soybean, and winter wheat) models submitted data (Figure S1) so the model ensemble standard deviation can be calculated uniformly in each case. JULES did not simulate winter wheat and LPJ-GUESS did not simulate rice and soybean. Emulator performance is generally satisfactory, with some exceptions. Emulator failures (significant areas of poor performance) occur for individual crop-model combinations, with performance generally degrading for colder and wetter scenarios.

emulator criterion is that its errors be small relative to inter-model differences. The normalized error $e$ is defined separately for each C,T,W,N scenario $s$ as the difference between emulated and simulated fractional yield changes, normalized by the standard deviation in simulated changes across all models:

$$e_s = \frac{F_{em,\,s} - F_{sim,\,s}}{\sigma_{sim,\,s}} \tag{4}$$

where $F$ is the fractional change in yields $Y$ between scenario $s$ and baseline $b$:

$$F_s = \frac{Y_s - Y_b}{Y_b} \tag{5}$$

We calculate the mean error for each grid cell, model, and crop in each C,T,W,N scenario by comparing emulated and simulated yields. A normalized error $e < 1$ means that any deviation of the emulation from the simulation is less than 1 standard deviation of the inter-model spread.

Evaluation of this metric implies that GGCMI Phase~~II~~ 2 emulators are generally satisfactory. Emulator performance is illustrated in Figure 7, which shows all models and crops crops over currently cultivated area. Over all crops and models,

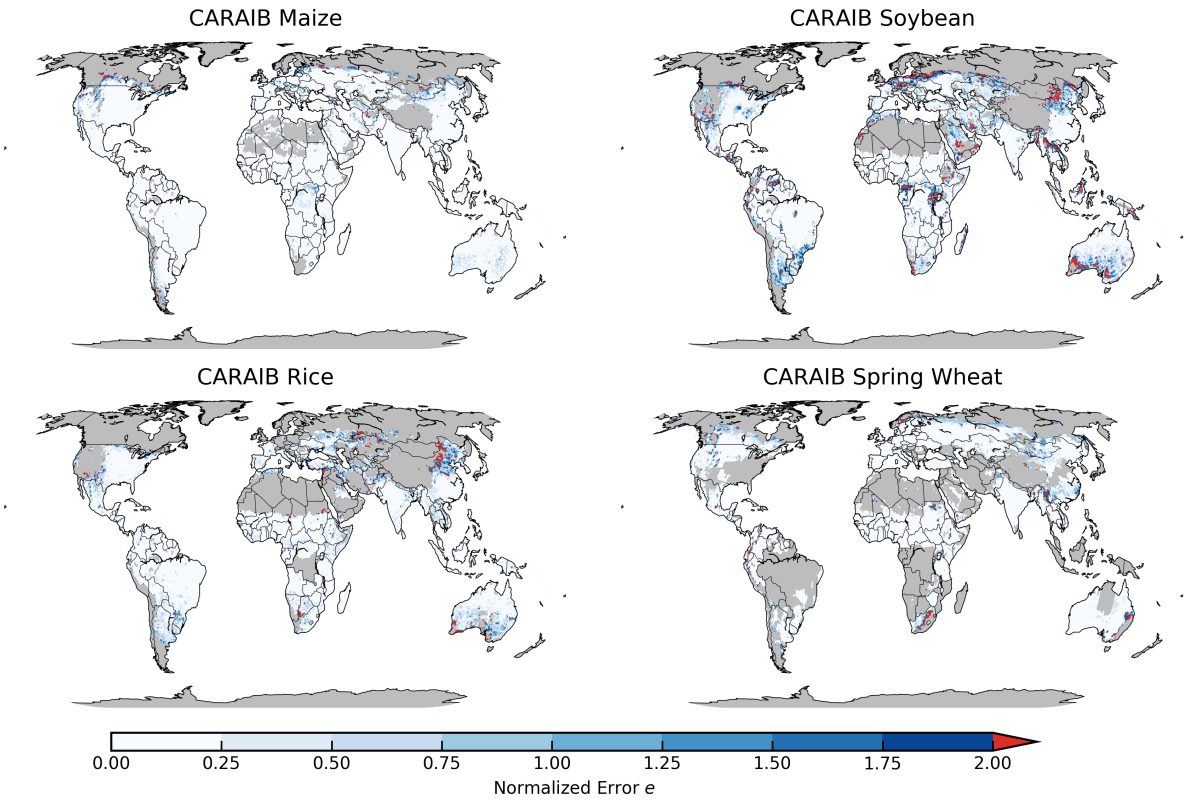


**Figure 8.** Illustration of our first test of emulator performance, applied to the CARAIB model for the T+4 scenario for rainfed crops. Colors indicate the normalized emulator error $e$, where $e > 1$ means that emulator error exceeds the multi-model standard deviation. For consistency, we show $e$ only for geographic areas simulated by at least six models and where baseline yields are greater than 0.5 ton ha$^{-1}$. Emulator performance is generally good relative to model spread in areas where crops are currently cultivated (compare to Figure S2-S3) and in temperate zones in general; emulation issues occur primarily in marginal areas with low yield potentials.

the average normalized error $e < 1$ over 95% of currently cultivated area. For maize, the most tractable crop to emulate, all 9 models return $e < 1$ over 97% of currently cultivated area. Only three crop-model combinations are problematic, returning $e < 1$ over less than 90% of cultivated area even when using the 34-term statistical model: PROMET and CARAIB for soybeans (79% and 83%), and JULES for spring wheat (85%). Misfits typically occur when models show strong discontinuities in yield response (as shown in Figure 5), or when carbon dioxide fertilization gains interact nonlinearly with changes in temperature or water. Including higher-order C terms helps in the latter case but does not reduce emulator errors to zero. See Supplemental Figures S22-S23 for examples of worst-case emulator failures.

While Figure 7 shows only currently cultivated land, performance can be worse in locations where crops are not currently cultivated, or on marginal lands where current potential yields are low. ~~(~~In general, emulator performance is poor anywhere that models show steep yield changes once some threshold has been reached~~, whether these are abrupt gains or~~ . Some of these case involve complete crop failures ~~.)~~ in a changed climate, but most involve yield improvements: abrupt gains in regions

that are too cold or dry under current conditions but that become viable given warming or wetting. Figure 8 illustrates this effect for CARAIB in the T+4 scenario, showing normalized error over all simulated area with non-zero baseline yield and at least 6 models providing simulations. CARAIB emulator performance is generally good where crops are grown but can be poor ($e > 2$) in arid or mountainous zones, e.g. the edges of the Sahara, Inner Mongolia, South Africa and Southern Australia. Effects will vary by crop model as they differ in process implementations; see the different model description papers referenced in Table 1 for more details. Note that the choice of statistical model for emulation involves a trade-off in the spatial pattern of errors~~: adding terms to the statistical model increases~~ . The 34-term statistical model used here maximally improves emulator fidelity in problematic "fringe" areas~~where crops are currently not cultivated, but reduces~~ , at the expense of lowering it slightly over high-yield areas. For example, over currently cultivated land, CARAIB maize emulators have normalized error $e < 1$ over ~~98.8% of currently cultivated land with the reduced 23-term Equation 2 but only 98.5~~98.5% of area with the full 34-term Equation 1 but over 98.8% with the ~~34-term Equation 1. Over simulated~~ reduced 23-term Equation 2. The effect is reversed over uncultivated land, with CARAIB maize emulators ~~have~~ showing $e < 1$ ~~for only 88.7~~over 93.7% of area with the ~~reduced Equation 2 but 93.7% with the~~ full Equation ~~1~~1 but only over 88.7% of area with the reduced Equation 2.

~~Note that the~~ The normalized error assessment is relatively forgiving for several reasons. First, it is an in-sample validation, with the emulation evaluated against the simulations actually used to train the emulator. Had we used a spline interpolation, the error would necessarily be zero. Second, the metric scales emulator fidelity not by the magnitude of yield changes in the evaluated model but by the spread in yield changes across models. The normalized error $e$ for a given model then depends on the particular suite of ~~other~~ models considered in the intercomparison exercise. The rationale for the choice is to relate the fidelity of the emulation to the true uncertainty, which we take as the multi-model spread, but the metric then has the property that where models differ more widely, the standard for emulators becomes less stringent, and vice versa. In GGCMI Phase~~II~~ 2 the effect is manifested in the higher normalized errors for soybeans across all models, which result not because soybean yields are difficult to emulate but because models agree more closely on yield changes for soybeans than for the other crops.

2. *Out-of-sample validation.* We provide a second, more stringent test of emulator performance via a 3-fold cross validation (also termed an out-of-sample validation). In this test the GGCMI Phase~~II~~ 2 dataset is split randomly into two parts, with 90% of the data used to train (calibrate) the model and the held-out 10% used to test (evaluate) the fidelity of the resulting emulator. ~~We~~ The procedure is repeated three times; in each case we calculate the root mean square error (RMSE) between the emulated (predicted) and actual simulated ~~values across the test set , repeat the process twice, and average the results of the two splits.~~ test set values, and then average the three results. The result is a single metric for each grid cell for each model-crop combination. As a last step, we normalize the ~~RMSE in~~ error metric for each grid cell by dividing by ~~the simulated yield change .~~ its maximum yield change over the entire CTWN dataset. (Since all models have submitted the extreme T+6 scenario, this normalization choice is not problematic.) Note that this validation exercise is independent of the procedure for generating the final published emulator values, which are generated using the full CTWN dataset.

The resulting error metric is generally low. Table 3 shows the yield-change-normalized RMSE for rainfed crops in all models over currently cultivated land, both in selected major producing regions and in the global average. ~~(~~We include all simulations in the CTWN space and ~~take~~ report the average error value ~~.) Mean~~ in Table 3. Global mean grid cell RMSE is below 5% of

maximum yield changes in all cases, or in absolute terms less than 0.2 ton ha$^{-1}$ for all except JULES ~~soy, which is~~ soybean simulations (0.36 ton ha$^{-1}$ ~~in the global mean. For irrigated crops , absolute emulator errorsare generally lower~~). Emulators for rainfed and irrigated crops have similar fractional errors, but since irrigated crops experience lower yield changes ~~the fractional errorsare similar~~across the CTWN scenarios, they also have lower absolute errors. See Supplemental Material Section 9 for maps of cross validation RMSE for each crop and model.

**Table 3.** RMSE of emulator replication of simulated yields of rainfed crops, stated as a percentage of simulated yield change. Values are the mean grid cell error as a percentage of simulated yield change, over all currently cultivated grid cells weighted by cultivation area, for selected major regions (NA: North America, SA: South America). For comparison, global mean values are show in parentheses. Errors are calculated using the 90-10 cross validation scheme described in text, with the model trained on 90% of the data and validated on the held-out 10% (repeated twice). All fits are made with the Bayesian Ridge method; for context we mark with * those cases where the Bayesian Ridge is required because the OLS linear model fails (e.g. PEPIC, which has the lowest number of samples at n=121).

| **Model** | **NA Maize~~%~~** | **SA Soybean~~%~~** | **SE Asian Rice~~%~~** | **NA S. Wheat~~%~~** | **European W. Wheat~~%~~** |
|---|---|---|---|---|---|
| **CARAIB** | 0.7 (0.9) | 2.4 (2.4) | 2.4 (2.4) | 1.3 (1.4) | 2.7 (1.9) |
| **EPIC-TAMU** | 2.4 (1.8) | 1.8 (2.6) | 1.6 (1.6) | 1.8 (1.9)* | 1.1 (1.1) |
| **JULES** | 2.6 (2.6) | 4.6 (4.0) | 1.6 (1.7) | 2.0 (2.2) | NA |
| **GEPIC** | 2.1 (2.4) | 1.0 (1.2) | 2.0 (2.1) | 3.7 (3.3) | 4.0 (2.9) |
| **LPJ-GUESS** | 1.0 (1.1) | NA | NA | 1.0 (1.3) | 1.0 (1.2) |
| **LPJmL** | 1.8 (1.8) | 1.1 (1.3) | 1.2 (1.1) | 0.8 (1.1) | 1.5 (1.3) |
| **pDSSAT** | 1.9 (1.7) | 1.2 (1.1) | 1.7 (1.6) | 1.1 (1.3) | 1.4 (1.5) |
| **PROMET** | 3.4 (2.7)* | 2.0 (2.7)* | 2.1 (1.8)* | 4.3 (3.7)* | 4.6 (3.4)* |
| **PEPIC** | 1.8 (1.8)* | 1.4 (1.9)* | 1.4 (1.4)* | 2.3 (2.3)* | 4.9 (2.9)* |

Note that this ~~metric is relatively simple and~~ relatively simple metric may be over-conservative. The randomized sampling protocol for dividing training and test sets can mean that a training set omits edge simulations at the highest or lowest value in CTWN space. The test prediction then involves extrapolating out of the training set range (e.g. predicting a T+6 case when the training set extends only to T+4), an improper use of an emulator. ~~Values~~ RMSE values would be lower ~~under a different sampling strategy~~ if we had used a more careful sampling strategy that precluded extrapolation (e.g. "leave-one-out"). For additional discussion of more detailed potential evaluation metrics, see e.g. Castruccio et al. (2014).

### 4.3 Emulation of realistic climate projections

Finally, we test the ability of an emulator based on the GGCMI Phase~~II~~ 2 perturbed mean training set to reproduce the response of a crop model driven by a realistic evolving climate scenario. ~~The goal is~~ Our emulators are trained only on growing-season means, and the GGCMI Phase 2 exercise involved only changes in means. We therefore seek to assess whether ~~effects of future changes in~~ changes in the higher moments of temperature and precipitation distributions ~~are strong enough to~~ in a climate projection might have effects that lead to significant emulator error. Note that we are not asking whether year-over-year climate

variability matters to crop yields; this point is well-established (Ray et al., 2015). The question instead is whether a realistic future climate projections involves *changes* in variability large enough that they compromise an emulator based on the GGCMI PhaseII dataset. We first drive 2 dataset.

To assess this potential error, we generate new crop model simulations using the LPJmL crop model (taken as a representative of GGCMI models)with climate model output under the high-end RCP 8.5 scenario. We choose for this purpose , driven by a climate simulation from the the Coupled Model Intercomparison Project Phase 5 (CMIP5) archive (Jones et al., 2011; Martin et al., 2011; T . To maximize any potential bias, we choose a climate model (HadGEM2-ES) with that exhibits relatively large changes in growing-season temperature variability among CMIP5 members (Supplemental Table S1)among members of the Coupled Model Intercomparison Project Phase 5 (CMIP-5) archive (Jones et al., 2011; Martin et al., 2011; Taylor et al., 2012), and use the high-end RCP 8.5 scenario. We also hold $CO_2$ fixed to emphasize the results of temperature and precipitation changes, in the absence of the beneficial effects of increased $CO_2$. We then drive the LPJmL emulator with compare the resulting simulated yields to the output of the GGCMI LPJmL emulator driven by the HadGEM2-ES yearly growing season anomalies, and evaluate how well the resulting emulated yields reproduce those simulated under the full climate scenario. The comparison suggests yearly growing-season T and P anomalies (Figure 9). The GGCMI LPJmL emulator is able to capture the yield changes well: for all crops, emulated and simulated global production in the last decade of the simulation are identical to within 1.5%. These results imply that globally, the results of future distributional shifts on climatological yields are small relative to the effects of mean changes (Figure 9). In the LPJmL example of Figure 9, emulated and simulated global production in the last decade of the simulation are identical to within 1.5% for all crops. Emulators The GGCMI LPJmL emulators also reproduce decadal variations in yields, which are especially strong in spring wheat grown in northern latitudes , (Figure 9, right), and even capture much of the residual year-to-year yield variability. (: $R^2$ of emulated vs. simulated annual yield anomalies relative to the 10-year running mean is 0.8 for spring wheat (and $\sim$0.3 for all other crops.)).

Distributional effects might be expected to be stronger at high latitudes, because temperature and precipitation variability are larger there, so that changes in variability can be correspondingly more important. However, We find however that most crops (spring wheat, winter wheat, and maize) show no emulator bias that grows with latitude. Rice is the exception: the climatological-mean emulator slightly over-predicts yield losses in the tropics and under-predicts losses at higher latitudes (where little rice is currently grown). Poleward of 30 degrees latitude, the LPJmL simulation under the HadGEM2 RCP scenario shows a 49% reduction in rice yields by end-of-century (without growing-season adaptation), but the GGCMI-based emulator produces a reduction of only 39% (Supplemental Figure S11). These losses are concentrated in the lower mid-latitudes: only 21% of global rice is cultivated poleward of 30 degrees, and only 1% poleward of 45 degrees.

It is worth noting two complications involved in comparing emulated to simulated yields under a realistic climate change scenario, as in Figure 9. First, it is not trivial to choose how to relate temperature or precipitation in the evolving climate scenario to the $T$ and $P$ offsets used as regressors to the emulator. Using growing-season mean temperature can lead to complications if crop models assume that growing season lengths shift under climate change. For consistency, we match the temperature changes in the climate scenario to their equivalent emulator regressors by calculating means over the fixed baseline growing season. This choice ensures that the emulation is appropriately matched to the simulation. Second, while the emulator outputs

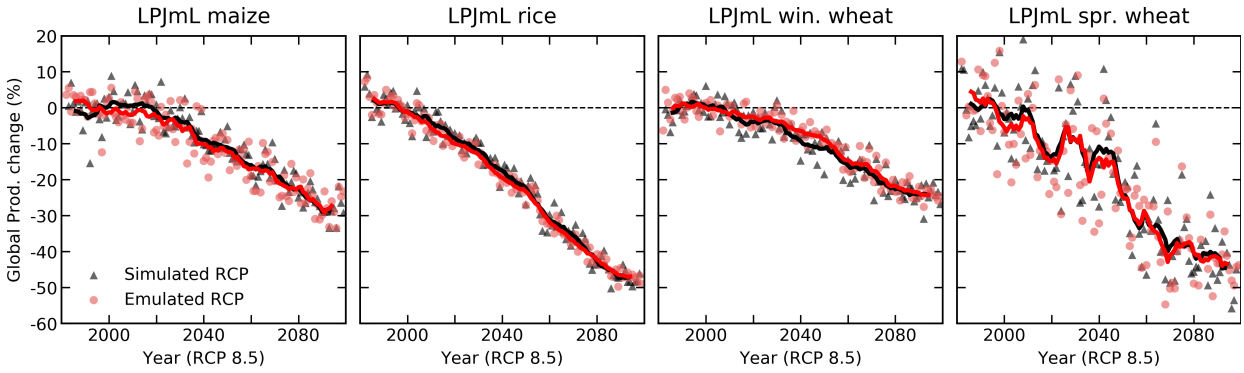


**Figure 9.** Test of emulator performance in reproducing yield simulations made with a realistic climate projection. Panels show simulated (black) and emulated (red) global production for four crops from the LPJmL model, driven with temperature and precipitation outputs from the HadGEM2-ES climate model for the RCP8.5 scenario. In both cases nitrogen and $CO_2$ are held fixed, at 200 kg $ha^{-1}$ and 360 ppm. Points show yearly global production change from the 1981-2010 baseline, and lines show a 10-year running mean. See text for discussion of relating the HadGEM2-ES temperature timeseries to the appropriate offset used in emulation. Emulators trained on uniform climatological offsets reproduce well the simulated production response under a realistic climate scenario: yields at end of century match to within 1.5%.

an estimated yield change, the baseline from which that yield change is calculated will be different between simulation and emulation, because the historical climate timeseries are not identical. For example, the baseline (1981-2010) yield of winter wheat simulated by LPJmL using the AgMERRA timeseries as part of GGCMI Phase~~II~~ 2 is 7% lower than that simulated using the HadGEM2-ES timeseries. To minimize the effects of different historical climate assumptions, we drive the emulator with the anomaly of the climate scenario from its own 1981-2010 mean. Bias in the historical climate timeseries could in theory produce discrepancies between emulated and simulated yield changes because of the nonlinearities discussed in Section 2.2, but the effect appears to play little role in the LPJmL comparison of Figure 9.

## 5 Emulator results and products

The crop model emulators developed here can be used for a variety of applications, because the emulator transforms the discrete simulation samples into a continuous response surface at any geographic scale. One use is construction of continuous agricultural damage functions in a flexible format. As an example, we present in Figure 10 global damage functions over each of the four dimensions tested in this study, constructed from the 4D emulation of each crop model.

These damage functions are useful in diagnosing commonalities and differences in the responses of crop models. In most cases, models agree on the sign of responses to individual factors, but the spread in model responses is comparable to the median response. Inter-model spreads are largest for spring wheat and smallest for soybeans, as also shown in Figure 7. Model responses to individual factors conform to expectations. As expected, the $CO_2$ response is smallest for maize, which is a C4 ~~grass~~crop, and the nitrogen response is smallest for soybeans, which are efficient fixers of atmospheric nitrogen. Nitrogen responses in crops other than soybeans are relatively similar, and most models show saturation beginning at values less than 200

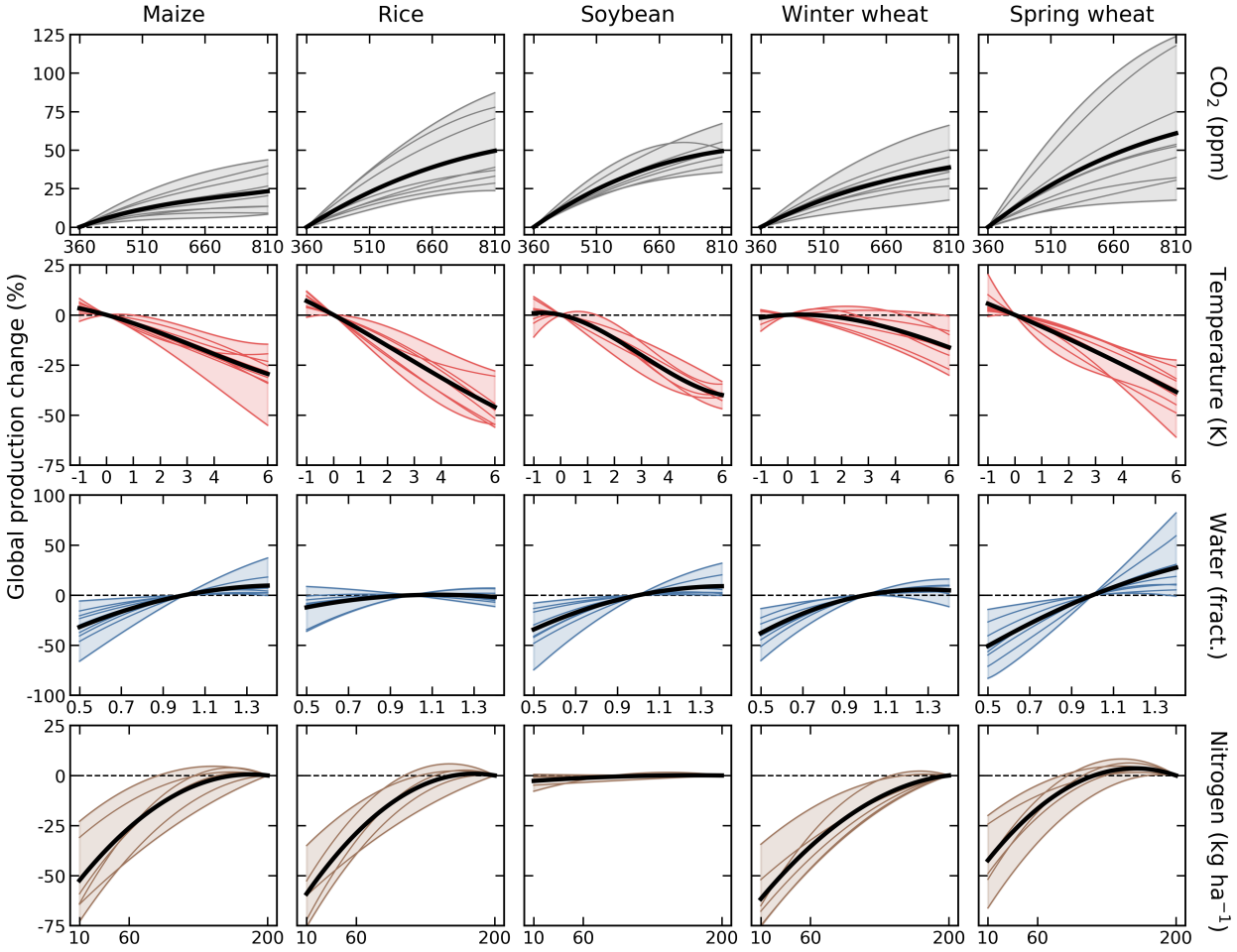


**Figure 10.** Emulated global damage functions for the five ~~GCCMI Phase II~~ crops over the four CTWN dimensions varied ~~.~~ in GGCMI Phase 2. Black line shows the multi-model mean and shaded area and colored lines the individual models. The number of models in each case varies, because some models did not provide all crops or simulate the N dimension. Each panel shows response to one covariate for rainfed crops, with all others held constant at baseline values (e.g. C = 360 ppm, N = 200 kg $ha^{-1}$). Damages are reported as percent change in global production over currently cultivated land relative to the 1981-2010 baseline. Note that y-axis ranges are not uniform. As expected, the N response is smallest in soybeans, which are nitrogen fixers, and the C response smallest in maize, which is a C4 crop. See Supplemental Figure S12 for an analogous figure identifying each crop model, and Supplemental Figure S13 for damage functions for the A1 (~~adpative~~ adaptive growing season) emulators, which have reduced temperature responses.

kg $ha^{-1}$. In nearly all crop models and for all crops except spring wheat, damages from reduced precipitation exceed benefits from increased precipitation. Spring wheat is the exception, likely because it is grown in high latitudes where rainfall may be limiting. Rice, by contrast, which is generally grown in locations with abundant water, shows nearly no benefit from increased

precipitation. Note that these damage functions do not consider whether increased precipitation might permit cultivation in new areas, and also that crop models generally do represent damages from excess soil moisture well (Li et al., 2019).

The GGCMI PhaseII 2 emulators are also intended as a tool for impacts assessments. The T and W functions presented in Figure 10 are not true global projections, because they emulate the consequences of uniform shifts across the globe. However, the emulator allows building analogous damage functions based on climate model output, which has more realistic spatial patterns of changes in temperature and precipitation. In Figure 11, we show emulated maize responses for 3 crop models under the RCP8.5 scenario, using output from 5 3 climate models from the CMIP 5 CMIP5 archive. Losses are shown as a function of mean growing-season temperature over currently cultivated land. While these damages functions aggregate over all currently cultivated land, the global coverage of GGCMI PhaseII 2 allows impacts modelers to develop damage functions for any desired geopolitical or geographic region larger than 0.5 degrees in latitude and longitude.

The emulated responses of Figure 11 allow diagnosing the factors of greatest importance to projected yield changes under future climate change. In the maize example here, temperature is the overwhelmingly dominant factor for pDSSAT, but $CO_2$ responses are far larger in PROMET. (CO2 is important across models for spring wheat, see Figure S14. ) For all crop models, the aggregated effects of precipitation changes are negative, exacerbating yield losses (compare T and T+W cases), because precipitation in HadGEM2 actually declines over maize cultivation regions, especially in Central and S. America. Precipitation effects are relatively small, however, as manifested in two ways: as only a small mean shift in yield projections for individual crop models (compare T and T+W cases), and as a relatively small increase in the spread of points here at a given temperature, despite the fact that the climate projections used involve different relationships between temperature and precipitation change. By contrast, the carbon dioxide fertilization response for PROMET is so large that projections from climate models of different sensitivities ($\Delta T/\Delta CO_2$) become clearly separated in Figure 11. PROMET yield responses would be more similar if plotted as a function of $CO_2$ than they are when plotted as in Figure 11 as a function of temperature change.

Disaggregating the factors driving crop yield changes also highlights the fact that errors of emulation are much smaller than the spread across crop models or even across different climate simulations. PROMET is the most quantitatively difficult model to emulate for maize, but its comparatively large emulation error (compare open squares to lines in T case) is still smaller than the spread simply due to different T patterns across climate simulations (Figure 11, left, compare differences between open squares and line with the spread in circles dots for a given temperature value). Uncertainties in the yield damage function due to projected patterns of temperature change are in turn smaller than spread due to differing model relationships of W and T changes (Figure 11, middle), and for PROMET are enormously outweighed by uncertainty in climate sensitivity (Figure 11, right). While emulator fidelity is important to ensure, it is important to recognize that these other uncertainties will dominate any impacts assessment exercise. Note that the pattern-related yield effects are actually relatively small for maize. (In Figure 11, left, compare lines, which show yield changes under uniform temperature shifts, to circlesdots, which show changes under realistic warming scenarios). Pattern-related yield effects can be larger for other crops, and the uncertainties due to climate projection differences correspondingly larger: see for example soybeans in Supplemental Figure S15.

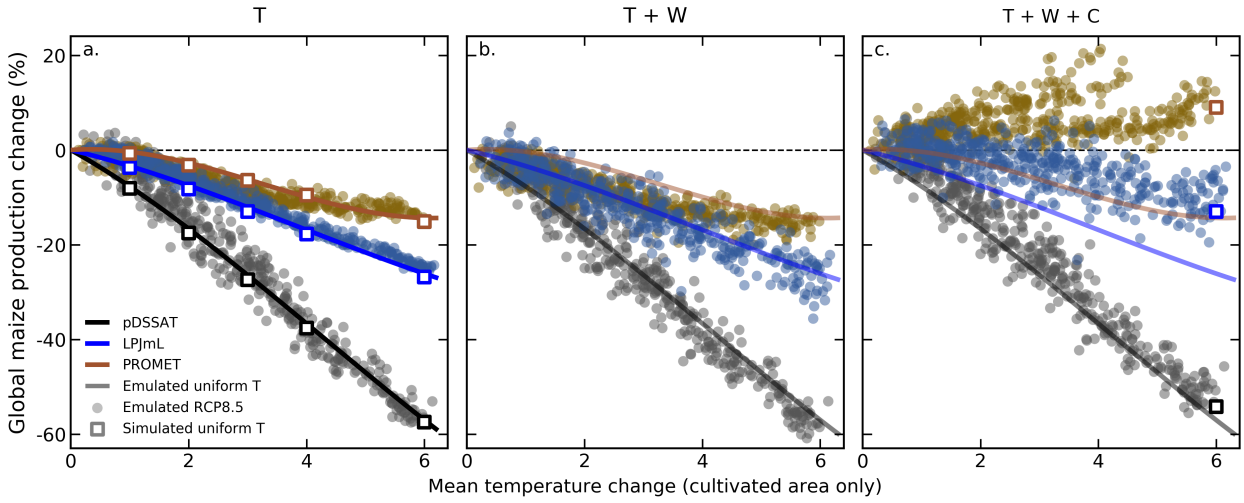


**Figure 11.** Illustration of the use of the emulator to study the factors affecting yields in more realistic climate ~~scenarios, for three different crop models~~scenario. Figure shows emulated yield changes (relative to 1981-2010) for maize (both rainfed and irrigated) on currently cultivated land under RCP8.5 climate projections from ~~5~~ 3 representative ~~CMIP-5~~ CMIP5 climate models ((HadGEM2-ES, GFDL-ESM2M, and IPSL-CM5A-LR), using changes to T only (**a**), to T and W (**b**), and to T, W, and C (**c**). ~~X-axis~~ The x-axis is the mean growing-season temperature change over cultivated land, computed using the historical growing season; note that these values will be higher than the corresponding global mean temperature change. ~~Circles~~ Dots are emulated yearly global production changes to 2100 (90 years $\times$ 5 climate timeseries = 450 per crop model), with x-axis the mean historical growing-season T shift over all grid cells where maize is grown (unweighted by within-cell cultivated area). **a:** Using only temperature changes allows comparing regional simulated and emulated values. Open squares are GGCMI Phase 2 simulated values for each T level, with CWN held at baseline; bold lines are emulated values over uniform $\Delta$T shifts (repeated in each panel). Emulation uncertainty (compare squares to lines) is small relative to differences across climate and crop models, and mean yield changes are similar whether T changes are applied as a uniform shift or in a more realistic spatial pattern (compare lines to ~~circles~~dots). **b:** Adding in precipitation changes increases yield spread across climate projections and depresses yield slightly. No squares are shown in **b** because the GGCMI uniform offsets of both T and W are not directly comparable to GCM-specific changes of T and W in a climate projection. **c:** $CO_2$ fertilization is small in pDSSAT, moderate in LPJmL, and very large in PROMET. The separation of groups of points in PROMET (gold) results because ~~CMIP-5~~ CMIP5 climate sensitivities differ by nearly a factor of two; points at far right are under the highest-sensitivity model, HadGEM2-ES. In RCP8.5, the 30-year-average $CO_2$ at end of century is 807 ppm (Riahi et al., 2011). For comparison, open squares in **c** show ~~GGCMI-II~~ GGCMI-2 simulated production changes at T+6, W=0, C=810 ppm. (Note that in these climate projections, mean $CO_2$ levels when T $>$ 5.8 degrees is 912 ppm.) See Supplemental Figures S14–15 for analogous figures for other crops (spring wheat and soybeans).

## 6 Discussion and conclusions

490 In this work we describe a new class of global gridded crop model emulators for 5 crops (maize, ~~soy~~soybean, rice and spring and winter wheat) and 9 process-based crop models, based on the GGCMI Phase~~II dataset. The systematic parameter sampling of the GGCMI Phase II experiment allows emulating~~ 2 dataset, a set of crop model simulations run with systematic perturbations

to carbon, temperature, precipitation, and nitrogen (CTWN). The goal of this project is to provide a lightweight tool that reproduces the output of large numerical simulations of process-based crop models. The resulting emulators should provide useful tools both for diagnosing crop model behavior and for climate impacts assessment, at least of large-scale time-averaged responses. Specific findings of this work include that:

- In crop models, the climatological mean yield responses to uniform perturbations in growing-season mean temperature and precipitation are very distinct from responses to historical weather fluctuations associated with the same mean differences. This result suggests that when emulating crop models, care must be taken if considering responses on both short and long timescales. The large GGCMI Phase 2 experiment allows us to emulate climatological-mean ~~crop yield~~ responses with a ~~relatively~~ simple statistical model without relying on the "natural experiment" of year-over-year variations.

- Climatological mean responses in all models can be well-fit with a simple third order polynomial in mean growing-season C, T, W, and N. The large GGCMI training set allows fitting in most cases with OLS, but use of a Bayesian Ridge regression provides additional stability and prevents overfitting. For most crop models, emulation is also possible with a simplified version of the statistical model with only 23 terms.

- The resulting emulators are highly flexible: they capture the strong geographic difference in crop yields and yields responses, can perform well on models with quite different sensitivities to climate or $CO_2$ changes. Emulators can faithfully reproduce the output of process-based crop models in both in- and ~~isolating long-term impacts from confounding factors that lead to different year-to-year responses. Across all models , emulation~~ out-of-sample tests. Emulation error is generally small other than in localized regions where crops are not currently grown: across all models and scenarios, errors over currently cultivated land never exceed 5% of yield changes at either global or regional scale. ~~The systematic sampling provides information on the influence of multiple interacting factors in a way that realistic climate model simulations cannot, and the use of a parametric statistical model allows physical interpretation of parameter values. While emulators based on the GGCMI PhaseII protocol of uniform perturbations to historical climate will not reproduce any effects of changing variability in future climate projections (any temperature variability changes or precipitation variabilitychanges other than multiplicative mean shifts) , in practice these effects appear to be small~~

- Emulators trained on the GGCMI Phase 2 dataset, which samples over uniform climate perturbations, can effectively reproduce the behavior of crop models driven by realistic future projections of future T and P changes. This result suggests that any projected changes in weather distributions (temperature and precipitation variability) have relatively little effect on crop model yield responses relative to changes in means, at least on the regionally aggregated level. ~~Emulators provide~~

- The GGCMI emulators should provide a powerful tools for both model comparison and impacts assessments~~by capturing the responses of~~ . The emulators can be used to develop standalone damage functions at any geographic scale larger that 0.5 degrees, or can be integrated directly into a larger integrated assessment model (IAM) framework. Emulators can

also be used to study differences across crop models in responses to individual drivers of yield changes, making them useful for model comparison and improvement.

While an emulator that captures the response of a process-based crop ~~models~~ model in a lightweight form ~~. The emulators provide over~~ will never be more *accurate* then its parent model, it can have multiple advantages over a numerical simulation. Emulation over the systematic sampling of the GGCMI Phase 2 experiment provides information on the influence of multiple interacting factors in a way that individual, more realistic process-based model runs cannot. Because we use a parametric statistical model, fitted parameter values can be physical interpreted to help understand differences between crop models. The flexibility and low computational requirements of emulators also make them particularly suitable for applications in integrated climate change impact assessments and projections of land-use change (e.g. Nelson et al., 2014a). Data storage requirements are reduced by three orders of magnitude~~reduction in data storage~~: the yield output for a single crop model ~~that simulates all GGCMIP PhaseII~~ simulating all GGCMI Phase 2 scenarios for 5 crops is ∼12.5 GB~~;~~, while the equivalent global gridded emulator parameters are only ∼20 MB~~and allow emulation of arbitrary future scenarios~~. Computational requirements are nearly negligible: a thousand years of global 0.5 degree yields, i.e ∼40,000,000 individual yield projections, can be emulated in 20 seconds on a laptop computer. The ~~emulators can be used to develop standalone damage functions at any geographic scale larger that 0.5 degrees, or can be integrated directly into a larger integrated assessment model (IAM) framework~~resulting suite of emulators should find considerable use in climate impacts analyses (e.g. Stevanović et al., 2016), and allow explicit evaluation of the uncertainty embedded in the choice of climate and crop models (Müller et al., 2017).

Several cautions should be noted when using the emulators presented here. First, extrapolation outside the GGCMI Phase~~II~~ 2 sample space should be avoided. Polynomial fits, while faithful within sample, quickly become non-physical outside of the tested range. This constraint is important given the strong warming expected under high-end greenhouse gas concentration scenarios (e.g. RCP8.5): if growing seasons are held fixed, climate model projections yield mean temperatures changes above 6K by end of century in many agricultural regions. Second, while the emulators are valuable for understanding the shape of yield responses and the factors that drive them, the absolute values of emulated yields should be treated with caution. ~~Because the GGCMI PhaseII experiment was designed to focus on yield changes and not on replicating real-world yields, most~~ The GGCMI Phase 2 models are not formally calibrated ~~, and their~~ and so the emulators should be used for absolute ~~impacts~~ projections only in combination with historical ~~yield data. The GGCMI Phase II~~ data. Third, neither growing season specification tested in GGCMI Phase 2 (A0 and A1) accounts for a major potential adaptation pathway under climate change, a shift to earlier or later planting dates (Waha et al., 2012) or generally different growing seasons (Minoli et al., 2019a). And finally, the emulator should not be used to predict individual yearly yields, as the forced climatological mean yield response will not match the response to mean growing season weather in a single year. The emulator cannot provide a measure of changing yield variance, and should not be used to evaluate extremes.

In summary, the GGCMI Phase 2 dataset and emulators invite a broad range of potential future avenues of analysis. Future studies using the emulators described here could include a detailed examination of interaction terms, robust quantification of model sensitivities to input drivers, and evaluation of geographic shifts in optimal growing regions. The large suite of crop models emulated lends itself particularly well to model comparison efforts, including identifying locations of model consensus

(or lack thereof) and causes of model differences. Studies of yield responses to changes in growing-season variability would require new simulations, but the emulators presented here provide a ready means of testing the null hypothesis that such effects are small. Similar structured training sets could be constructed to directly study responses to variability changes: see e.g. Poppick et al. (2016); Haugen et al. (2018) for methods of constructing synthetic climate timeseries with altered variability. The GGCMI PhaseH 2 dataset can be used as a testbed for examining the ability of statistical models using more detailed within-season regressors to capture both year-over-year and climatological changes, and for more systematic studies of emulation itself, including evaluation of alternate statistical specifications or machine learning methods. In general, the GGCMI PhaseH 2 experiment demonstrates the promise and utility of systematic parameter sweeps for improving understanding of the factors driving crop responses and for evaluating and improving process-based crop models.

*Code and data availability.* The polynomial parameters for crop model emulators are available at https://doi.org/10.5281/zenodo.3592453.

*Author contributions.* J.E., C.M, A.R., J.F., and E.M. designed the research. C.M., J.J., P.F., C.F., L.F., R.C.I., I.J., C.J., W.L., S.O., M.P., T.P., A.Re., K.W., and F.Z. performed the simulations. J.F., J.J., A.S., M.L., Z.W., and E.M. performed the analysis and J.F., C.M., and E.M. prepared the manuscript. All authors contibuted to editing the manuscript.

*Competing interests.* The authors declare no competing interests.

*Acknowledgements.* We thank Michael Steinand , Kevin Schwarzwald, and three anonymous reviewers, who provided helpful suggestions that contributed to this work. This research was performed as part of the Center for Robust Decision-making on Climate and Energy Policy (RDCEP) at the University of Chicago, and was supported through a variety of sources. RDCEP is funded by NSF grant #SES-1463644 through the Decision Making Under Uncertainty program. J.F. was supported by the NSF NRT program, grant #DGE-1735359 and by an NSF Graduate Research Fellowship, grant #DGE-1746045. C.M. was supported by the MACMIT project (01LN1317A) funded through the German Federal Ministry of Education and Research (BMBF). C.F. was supported by the European Research Council Synergy grant #ERC-2013-SynG-610028 Imbalance-P. P.F. and K.W. were supported by the Newton Fund through the Met Office Climate Science for Service Partnership Brazil (CSSP Brazil). K.W. was supported by the IMPREX research project supported by the European Commission under the Horizon 2020 Framework programme, grant #641811. A.S. was supported by the Office of Science of the U.S. Department of Energy as part of the Multi-sector Dynamics Research Program Area. S.O. acknowledges support from the Swedish strong research areas BECC and MERGE together with support from LUCCI (Lund University Centre for studies of Carbon Cycle and Climate Interactions). R.C.I. acknowledges support from the Texas Agrilife Research and Extension, Texas A & M University. computing resources were provided by the University of Chicago Research Computing Center (RCC). This is paper number 36 of the Birmingham Institute of Forest Research. Computing resources were provided by the University of Chicago Research Computing Center (RCC).

~~*This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. (DGE-1746045). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.*~~

*Financial support.* RDCEP is funded by NSF through the Decision Making Under Uncertainty program (grant #SES-1463644). James Franke was supported by the NSF NRT program (grant no. DGE-1735359) and the NSF Graduate Research Fellowship Program (grant #DGE-1746045). Christoph Müller was supported by the MACMIT project (grant no. 01LN1317A) funded through the German Federal Ministry of Education and Research (BMBF). Alex C. Ruane was supported by NASA NNX16AK38G (INCA) and the NASA Earth Sciences Directorate/GISS Climate Impacts Group. Christian Folberth was supported by the European Research Council Synergy (grant no. ERC-2013-SynG-610028) Imbalance-P. Pete Falloon and Karina Williams were supported by the Newton Fund through the Met Office program Climate Science for Service Partnership Brazil (CSSP Brazil). Karina Williams was supported by the IMPREX research project supported by the European Commission under the Horizon 2020 Framework program (grant no. 641811). Stefan Olin acknowledges support from the Swedish strong research areas BECC and MERGE, together with support from LUCCI (Lund University Centre for studies of Carbon Cycle and Climate Interactions). R. Cesar Izaurralde acknowledges support from the Texas Agrilife Research and Extension, Texas A & M University. Abigail Snyder was supported by the Office of Science of the U.S. Department of Energy as part of the Multi-sector Dynamics Research Program Area.

## References


Aulakh, M. S. and Malhi, S. S.: Interactions of Nitrogen with Other Nutrients and Water: Effect on Crop Yield and Quality, Nutrient Use Efficiency, Carbon Sequestration, and Environmental Pollution, Advances in Agronomy, 86, 341 – 409, https://doi.org/10.1016/S0065-2113(05)86007-9, 2005.

Blanc, E.: Statistical emulators of maize, rice, soybean and wheat yields from global gridded crop models, Agricultural and Forest Meteorology, 236, 145 – 161, https://doi.org/10.1016/j.agrformet.2016.12.022, 2017.

Blanc, E. and Sultan, B.: Emulating maize yields from global gridded crop models using statistical estimates, Agricultural and Forest Meteorology, 214-215, 134 – 147, https://doi.org/10.1016/j.agrformet.2015.08.256, 2015.

Castruccio, S., McInerney, D. J., Stein, M. L., Liu Crouch, F., Jacob, R. L., and Moyer, E. J.: Statistical Emulation of Climate Model Projections Based on Precomputed GCM Runs, Journal of Climate, 27, 1829–1844, https://doi.org/10.1175/JCLI-D-13-00099.1, 2014.

Challinor, A., Wheeler, T., Craufurd, P., Slingo, J., and Grimes, D.: Design and optimisation of a large-area process-based model for annual crops, Agricultural and Forest Meteorology, 124, 99 – 120, https://doi.org/https://doi.org/10.1016/j.agrformet.2004.01.002, http://www.sciencedirect.com/science/article/pii/S0168192304000085, 2004.

Challinor, A., Watson, J., Lobell, D., Howden, S., Smith, D., and Chhetri, N.: A meta-analysis of crop yield under climate change and adaptation, Nature Climate Change, 4, 287 – 291, https://doi.org/10.1038/nclimate2153, 2014.

Conti, S., Gosling, J. P., Oakley, J. E., and O'Hagan, A.: Gaussian process emulation of dynamic computer codes, Biometrika, 96, 663–676, https://doi.org/10.1093/biomet/asp028, 2009.

Dee, D. P., Uppala, S., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, d. P., et al.: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, Quarterly Journal of the royal meteorological society, 137, 553–597, 2011.

Dury, M., Hambuckers, A., Warnant, P., Henrot, A., Favre, E., Ouberdous, M., and François, L.: Responses of European forest ecosystems to 21st century climate: assessing changes in interannual variability and fire intensity, iForest - Biogeosciences and Forestry, pp. 82–99, https://doi.org/10.3832/ifor0572-004, 2011.

Elliott, J., Kelly, D., Chryssanthacopoulos, J., Glotter, M., Jhunjhnuwala, K., Best, N., Wilde, M., and Foster, I.: The parallel system for integrating impact models and sectors (pSIMS), Environmental Modelling and Software, 62, 509–516, https://doi.org/10.1016/j.envsoft.2014.04.008, 2014.

Ferrise, R., Moriondo, M., and Bindi, M.: Probabilistic assessments of climate change impacts on durum wheat in the Mediterranean region, Natural Hazards and Earth System Sciences, 11, 1293–1302, https://doi.org/10.5194/nhess-11-1293-2011, 2011.

Folberth, C., Gaiser, T., Abbaspour, K. C., Schulin, R., and Yang, H.: Regionalization of a large-scale crop growth model for sub-Saharan Africa: Model setup, evaluation, and estimation of maize yields, Agriculture, Ecosystems & Environment, 151, 21 – 33, https://doi.org/10.1016/j.agee.2012.01.026, 2012.

Franke, J. A., Müller, C., Elliott, J., Ruane, A. C., Jägermeyr, J., Balkovic, J., Ciais, P., Dury, M., Falloon, P. D., Folberth, C., François, L., Hank, T., Hoffmann, M., Izaurralde, R. C., Jacquemin, I., Jones, C., Khabarov, N., Koch, M., Li, M., Liu, W., Olin, S., Phillips, M., Pugh, T. A. M., Reddy, A., Wang, X., Williams, K., Zabel, F., and Moyer, E. J.: The GGCMI Phase 2 experiment: global gridded crop model simulations under uniform changes in $CO_2$, temperature, water, and nitrogen levels (protocol version 1.0), Geoscientific Model Development, 13, 2315–2336, https://doi.org/10.5194/gmd-13-2315-2020, https://gmd.copernicus.org/articles/13/2315/2020/, 2020.

Frieler, K., Lange, S., Piontek, F., Reyer, C. P. O., Schewe, J., Warszawski, L., Zhao, F., Chini, L., Denvil, S., Emanuel, K., Geiger, T., Halladay, K., Hurtt, G., Mengel, M., Murakami, D., Ostberg, S., Popp, A., Riva, R., Stevanovic, M., Suzuki, T., Volkholz, J., Burke, E., Ciais, P., Ebi, K., Eddy, T. D., Elliott, J., Galbraith, E., Gosling, S. N., Hattermann, F., Hickler, T., Hinkel, J., Hof, C., Huber, V., Jägermeyr, J., Krysanova, V., Marcé, R., Müller Schmied, H., Mouratiadou, I., Pierson, D., Tittensor, D. P., Vautard, R., van Vliet, M., Biber, M. F., Betts, R. A., Bodirsky, B. L., Deryng, D., Frolking, S., Jones, C. D., Lotze, H. K., Lotze-Campen, H., Sahajpal, R., Thonicke, K., Tian, H., and Yamagata, Y.: Assessing the impacts of 1.5$^o$C global warming — Simulation protocol of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP2b), Geosci. Model Dev., 10, 4321–4345, https://doi.org/10.5194/gmd-10-4321-2017, 2017.

Fronzek, S., Pirttioja, N., Carter, T. R., Bindi, M., Hoffmann, H., Palosuo, T., Ruiz-Ramos, M., Tao, F., Trnka, M., Acutis, M., Asseng, S., Baranowski, P., Basso, B., Bodin, P., Buis, S., Cammarano, D., Deligios, P., Destain, M.-F., Dumont, B., Ewert, F., Ferrise, R., François, L., Gaiser, T., Hlavinka, P., Jacquemin, I., Kersebaum, K. C., Kollas, C., Krzyszczak, J., Lorite, I. J., Minet, J., Minguez, M. I., Montesino, M., Moriondo, M., Müller, C., Nendel, C., Öztürk, I., Perego, A., Rodríguez, A., Ruane, A. C., Ruget, F., Sanna, M., Semenov, M. A., Slawinski, C., Stratonovitch, P., Supit, I., Waha, K., Wang, E., Wu, L., Zhao, Z., and Rötter, R. P.: Classifying multi-model wheat yield impact response surfaces showing sensitivity to temperature and precipitation change, Agricultural Systems, 159, 209–224, https://doi.org/10.1016/j.agsy.2017.08.004, 2018.

Gadgil, S., Rao, P. S., and Rao, K. N.: Use of climate information for farm-level decision making: rainfed groundnut in southern India, Agricultural Systems, 74, 431 – 457, https://doi.org/10.1016/S0308-521X(02)00049-5, 2002.

Glotter, M., Elliott, J., McInerney, D., Best, N., Foster, I., and Moyer, E. J.: Evaluating the utility of dynamical downscaling in agricultural impacts projections, Proceedings of the National Academy of Sciences, 111, 8776–8781, https://doi.org/10.1073/pnas.1314787111, 2014.

Glotter, M., Moyer, E., Ruane, A., and Elliott, J.: Evaluating the Sensitivity of Agricultural Model Performance to Different Climate Inputs, Journal of Applied Meteorology and Climatology, 55, 151113145618 001, https://doi.org/10.1175/JAMC-D-15-0120.1, 2015.

Hank, T., Bach, H., and Mauser, W.: Using a Remote Sensing-Supported Hydro-Agroecological Model for Field-Scale Simulation of Heterogeneous Crop Growth and Yield: Application for Wheat in Central Europe, Remote Sensing, 7, 3934–3965, https://doi.org/10.3390/rs70403934, 2015.

Hansen, J. and Jones, J.: Scaling-up crop models for climate variability applications, Agricultural Systems, 65, 43 – 72, https://doi.org/10.1016/S0308-521X(00)00025-1, 2000.

Hasegawa, T., Fujimori, S., Havlík, P., Valin, H., Bodirsky, B. L., Doelman, J. C., Fellmann, T., Kyle, P., Koopman, J. F., Lotze-Campen, H., Mason-D'Croz, D., Ochi, Y., Domínguez, I. P., Stehfest, E., Sulser, T. B., Tabeau, A., Takahashi, K., Takakura, J., Hans van Meij and, W.-J. v. Z., Wiebe, K., and Witzke, P.: Risk of increased food insecurity under stringent global climate change mitigation policy, Nature Climate Change, 8, 699–703, 2018.

Haugen, M., Stein, M., Moyer, E., and Sriver, R.: Estimating changes in temperature distributions in a large ensemble of climate simulations using quantile regression, Journal of Climate, 31, 8573–8588, https://doi.org/10.1175/JCLI-D-17-0782.1, 2018.

He, W., Yang, J., Zhou, W., Drury, C., Yang, X., D. Reynolds, W., Wang, H., He, P., and Li, Z.-T.: Sensitivity analysis of crop yields, soil water contents and nitrogen leaching to precipitation, management practices and soil hydraulic properties in semi-arid and humid regions of Canada using the DSSAT model, Nutrient Cycling in Agroecosystems, 106, 201–215, https://doi.org/10.1007/s10705-016-9800-3, 2016.

Holden, P., Edwards, N., PH, G., Fraedrich, K., Lunkeit, F., E, K., Labriet, M., Kanudia, A., and F, B.: PLASIM-ENTSem v1.0: A spatiotemporal emulator of future climate change for impacts assessment, Geoscientific Model Development, 7, 433–451, https://doi.org/10.5194/gmd-7-433-2014, 2014.

Holzkämper, A., Calanca, P., and Fuhrer, J.: Statistical crop models: Predicting the effects of temperature and precipitation changes, Climate Research, 51, 11–21, https://doi.org/10.3354/cr01057, 2012.

Howden, S. and Crimp, S.: Assessing dangerous climate change impacts on Australia's wheat industry, Modelling and Simulation Society of Australia and New Zealand, pp. 505–511, https://doi.org/-, 2005.

Hsiang, S., Kopp, R., Jina, A., Rising, J., Delgado, M., Mohan, S., Rasmussen, D. J., Muir-Wood, R., Wilson, P., Oppenheimer, M., Larsen, K., and Houser, T.: Estimating economic damage from climate change in the United States, Science, 356, 1362–1369, https://doi.org/10.1126/science.aal4369, 2017.

Ingestad, T.: Nitrogen and Plant Growth; Maximum Efficiency of Nitrogen Fertilizers, Ambio, 6, 146–151, 1977.

Izaurralde, R., Williams, J., Mcgill, W., Rosenberg, N., and Quiroga Jakas, M.: Simulating soil C dynamics with EPIC: Model description and testing against long-term data, Ecological Modelling, 192, 362–384, https://doi.org/10.1016/j.ecolmodel.2005.07.010, 2006.

Jones, C. D., Hughes, J. K., Bellouin, N., Hardiman, S. C., Jones, G. S., Knight, J., Liddicoat, S., O'Connor, F. M., Andres, R. J., Bell, C., Boo, K.-O., Bozzo, A., Butchart, N., Cadule, P., Corbin, K. D., Doutriaux-Boucher, M., Friedlingstein, P., Gornall, J., Gray, L., Halloran, P. R., Hurtt, G., Ingram, W. J., Lamarque, J.-F., Law, R. M., Meinshausen, M., Osprey, S., Palin, E. J., Parsons Chini, L., Raddatz, T., Sanderson, M. G., Sellar, A. A., Schurer, A., Valdes, P., Wood, N., Woodward, S., Yoshioka, M., and Zerroukat, M.: The HadGEM2-ES implementation of CMIP5 centennial simulations, Geoscientific Model Development, 4, 543–570, https://doi.org/10.5194/gmd-4-543-2011, 2011.

Jones, J., Hoogenboom, G., Porter, C., Boote, K., Batchelor, W., Hunt, L., Wilkens, P., Singh, U., Gijsman, A., and Ritchie, J.: The DSSAT cropping system model, European Journal of Agronomy, 18, 235 – 265, https://doi.org/10.1016/S1161-0301(02)00107-7, 2003.

Li, Y., Guan, K., Schnitkey, G. D., DeLucia, E., and Peng, B.: Excessive rainfall leads to maize yield loss of a comparable magnitude to extreme drought in the United States, Global Change Biology, 25, 2325–2337, https://doi.org/10.1111/gcb.14628, 2019.

Lindeskog, M., Arneth, A., Bondeau, A., Waha, K., Seaquist, J., Olin, S., and Smith, B.: Implications of accounting for land use in simulations of ecosystem carbon cycling in Africa, Earth System Dynamics, 4, 385–407, https://doi.org/10.5194/esd-4-385-2013, 2013.

Liu, B., Asseng, S., Müller, C., Ewert, F., Elliott, J., Lobell, D. B., Martre, P., Ruane, A. C., Wallach, D., Jones, J. W., et al.: Similar estimates of temperature impacts on global wheat yield by three independent methods, Nature Climate Change, 6, 1130, 2016a.

Liu, J., Williams, J. R., Zehnder, A. J., and Yang, H.: GEPIC - modelling wheat yield and crop water productivity with high resolution on a global scale, Agricultural Systems, 94, 478 – 493, https://doi.org/10.1016/j.agsy.2006.11.019, 2007.

Liu, W., Yang, H., Folberth, C., Wang, X., Luo, Q., and Schulin, R.: Global investigation of impacts of PET methods on simulating crop-water relations for maize, Agricultural and Forest Meteorology, 221, 164 – 175, https://doi.org/10.1016/j.agrformet.2016.02.017, 2016b.

Liu, W., Yang, H., Liu, J., Azevedo, L. B., Wang, X., Xu, Z., Abbaspour, K. C., and Schulin, R.: Global assessment of nitrogen losses and trade-offs with yields from major crop cultivations, Science of The Total Environment, 572, 526 – 537, https://doi.org/10.1016/j.scitotenv.2016.08.093, 2016c.

Lobell, D. B. and Burke, M. B.: On the use of statistical models to predict crop yield responses to climate change, Agricultural and Forest Meteorology, 150, 1443 – 1452, https://doi.org/10.1016/j.agrformet.2010.07.008, 2010.

Lobell, D. B. and Field, C. B.: Global scale climate-crop yield relationships and the impacts of recent warming, Environmental Research Letters, 2, 014 002, https://doi.org/10.1088/1748-9326/2/1/014002, 2007.

MacKay, D.: Bayesian Interpolation, Neural Computation, 4, 415–447, https://doi.org/10.1162/neco.1992.4.3.415, 1991.

Makowski, D., Asseng, S., Ewert, F., Bassu, S., Durand, J., Martre, P., Adam, M., Aggarwal, P., Angulo, C., Baron, C., Basso, B., Bertuzzi, P., Biernath, C., Boogaard, H., Boote, K., Brisson, N., Cammarano, D., Challinor, A., Conijn, J., and Wolf, J.:

Statistical Analysis of Large Simulated Yield Datasets for Studying Climate Effects, p. 1100, World Scientific Publishing Co, https://doi.org/10.13140/RG.2.1.5173.8328, 2015.

Martin, T. H. D. T. G. M., Bellouin, N., Collins, W. J., Culverwell, I. D., Halloran, P. R., Hardiman, S. C., Hinton, T. J., Jones, C. D., McDonald, R. E., McLaren, A. J., O'Connor, F. M., Roberts, M. J., Rodriguez, J. M., Woodward, S., Best, M. J., Brooks, M. E., Brown, A. R., Butchart, N., Dearden, C., Derbyshire, S. H., Dharssi, I., Doutriaux-Boucher, M., Edwards, J. M., Falloon, P. D., Gedney, N., Gray, L. J., Hewitt, H. T., Hobson, M., Huddleston, M. R., Hughes, J., Ineson, S., Ingram, W. J., James, P. M., Johns, T. C., Johnson, C. E., Jones, A., Jones, C. P., Joshi, M. M., Keen, A. B., Liddicoat, S., Lock, A. P., Maidens, A. V., Manners, J. C., Milton, S. F., Rae, J. G. L., Ridley, J. K., Sellar, A., Senior, C. A., Totterdell, I. J., Verhoef, A., Vidale, P. L., and Wiltshire, A.: The HadGEM2 family of Met Office Unified Model climate configurations, Geoscientific Model Development, 4, 723–757, https://doi.org/10.5194/gmd-4-723-2011, 2011.

Mauser, W., Klepper, G., Zabel, F., Delzeit, R., Hank, T., Putzenlechner, B., and Calzadilla, A.: Global biomass production potentials exceed expected future demand without the need for cropland expansion, Nature Communications, 6, https://doi.org/10.1038/ncomms9946, 2015.

Minoli, S., Egli, D. B., Rolinski, S., and Müller, C.: Modelling cropping periods of grain crops at the global scale, Global and Planetary Change, 174, 35 – 46, https://doi.org/https://doi.org/10.1016/j.gloplacha.2018.12.013, http://www.sciencedirect.com/science/article/pii/S092181811830362X, 2019a.

Minoli, S., Müller, C., Elliott, J., Ruane, A. C., Jägermeyr, J., Zabel, F., Dury, M., Folberth, C., François, L., Hank, T., Jacquemin, I., Liu, W., Olin, S., and Pugh, T. A.: Global response patterns of major rainfed crops to adaptation by maintaining current growing periods and irrigation, Earth's Future, 0, https://doi.org/10.1029/2018EF001130, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018EF001130, 2019b.

Mistry, M. N., Wing, I. S., and De Cian, E.: Simulated vs. empirical weather responsiveness of crop yields: US evidence and implications for the agricultural impacts of climate change, Environmental Research Letters, 12, https://doi.org/10.1088/1748-9326/aa788c, 2017.

Moore, F. C., Baldos, U., Hertel, T., and Diaz, D.: New science of climate change impacts on agriculture implies higher social cost of carbon, Nature Communications, 8, https://doi.org/10.1038/s41467-017-01792-x, 2017.

Müller, C., Elliott, J., Chryssanthacopoulos, J., Arneth, A., Balkovic, J., Ciais, P., Deryng, D., Folberth, C., Glotter, M., Hoek, S., Iizumi, T., Izaurralde, R. C., Jones, C., Khabarov, N., Lawrence, P., Liu, W., Olin, S., Pugh, T. A. M., Ray, D. K., Reddy, A., Rosenzweig, C., Ruane, A. C., Sakurai, G., Schmid, E., Skalsky, R., Song, C. X., Wang, X., de Wit, A., and Yang, H.: Global gridded crop model evaluation: benchmarking, skills, deficiencies and implications, Geoscientific Model Development, 10, 1403–1422, https://doi.org/10.5194/gmd-10-1403-2017, 2017.

Müller, C., Elliott, J., Chryssanthacopoulos, J., Deryng, D., Folberth, C., Pugh, T. A. M., and Schmid, E.: Implications of climate mitigation for future agricultural production, Environmental Research Letters, 10, 125 004, https://doi.org/10.1088/1748-9326/10/12/125004, https://doi.org/10.1088%2F1748-9326%2F10%2F12%2F125004, 2015.

Nakamura, T., Osaki, M., Koike, T., Hanba, Y. T., Wada, E., and Tadano, T.: Effect of CO2 enrichment on carbon and nitrogen interaction in wheat and soybean, Soil Science and Plant Nutrition, 43, 789–798, https://doi.org/10.1080/00380768.1997.10414645, 1997.

Nelson, G. C., Mensbrugghe, D., Ahammad, H., Blanc, E., Calvin, K., Hasegawa, T., Havlik, P., Heyhoe, E., Kyle, P., Lotze-Campen, H., Lampe, M., Mason d'Croz, D., Meijl, H., Müller, C., Reilly, J., Robertson, R., Sands, R. D., Schmitz, C., Tabeau, A., Takahashi, K., Valin, H., and Willenbockel, D.: Agriculture and climate change in global scenarios: why don't the models agree, Agricultural Economics, 45, 85–101, https://doi.org/doi.org/10.1111/agec.12091, 2014a.

Nelson, G. C., Valin, H., Sands, R. D., Havlík, P., Ahammad, H., Deryng, D., Elliott, J., Fujimori, S., Hasegawa, T., Heyhoe, E., Kyle, P., von Lampe, M., Lotze-Campen, H., d'Croz, D. M., van Meijl, H., van der Mensbrugghe, D., Müller, C., Popp, A., Robertson, R., Robinson, S.,

Schmid, E., abd Andrzej Tabeau, C. S., and Willenbockel, D.: Climate change effects on agriculture: Economic responses to biophysical shocks, Proceedings of the National Academy of Sciences, 111, 3274–3279, 2014b.

O'Hagan, A.: Bayesian analysis of computer code outputs: A tutorial, Reliability Engineering & System Safety, 91, 1290 – 1300, https://doi.org/10.1016/j.ress.2005.11.025, 2006.

Olin, S., Schurgers, G., Lindeskog, M., Wårlind, D., Smith, B., Bodin, P., Holmér, J., and Arneth, A.: Modelling the response of yields and tissue C:N to changes in atmospheric $CO_2$ and N management in the main wheat regions of western Europe, Biogeosciences, 12, 2489–2515, https://doi.org/10.5194/bg-12-2489-2015, 2015.

Osaki, M., Shinano, T., and Tadano, T.: Carbon-nitrogen interaction in field crop production, Soil Science and Plant Nutrition, 38, 553–564, https://doi.org/10.1007/BF00025019, 1992.

Osborne, T., Gornall, J., Hooker, J., Williams, K., Wiltshire, A., Betts, R., and Wheeler, T.: JULES-crop: a parametrisation of crops in the Joint UK Land Environment Simulator, Geoscientific Model Development, 8, 1139–1155, https://doi.org/10.5194/gmd-8-1139-2015, 2015.

Ostberg, S., Schewe, J., Childers, K., and Frieler, K.: Changes in crop yields and their variability at different levels of global warming, Earth System Dynamics, 9, 479–496, https://doi.org/10.5194/esd-9-479-2018, 2018.

Oyebamiji, O. K., Edwards, N. R., Holden, P. B., Garthwaite, P. H., Schaphoff, S., and Gerten, D.: Emulating global climate change impacts on crop yields, Statistical Modelling, 15, 499–525, https://doi.org/10.1177/1471082X14568248, 2015.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825–2830, 2011.

Pirttioja, N., Carter, T., Fronzek, S., Bindi, M., Hoffmann, H., Palosuo, T., Ruiz-Ramos, M., Tao, F., Trnka, M., Acutis, M., Asseng, S., Baranowski, P., Basso, B., Bodin, P., Buis, S., Cammarano, D., Deligios, P., Destain, M., Dumont, B., Ewert, F., Ferrise, R., François, L., Gaiser, T., Hlavinka, P., Jacquemin, I., Kersebaum, K., Kollas, C., Krzyszczak, J., Lorite, I., Minet, J., Minguez, M., Montesino, M., Moriondo, M., Müller, C., Nendel, C., Öztürk, I., Perego, A., Rodríguez, A., Ruane, A., Ruget, F., Sanna, M., Semenov, M., Slawinski, C., Stratonovitch, P., Supit, I., Waha, K., Wang, E., Wu, L., Zhao, Z., and Rötter, R.: Temperature and precipitation effects on wheat yield across a European transect: a crop model ensemble analysis using impact response surfaces, Climate Research, 65, 87–105, https://doi.org/10.3354/cr01322, 2015.

Poppick, A., McInerney, D. J., Moyer, E. J., and Stein, M. L.: Temperatures in transient climates: Improved methods for simulations with evolving temporal covariances, Ann. Appl. Stat., 10, 477–505, https://doi.org/10.1214/16-AOAS903, 2016.

Portmann, F., Siebert, S., and Doell, P.: MIRCA2000 - Global Monthly Irrigated and Rainfed Crop Areas around the Year 2000: A New High-Resolution Data Set for Agricultural and Hydrological Modeling, Global Biogeochemical Cycles, 24, GB1011, https://doi.org/10.1029/2008GB003435, 2010.

Potter, N. J., Zhang, L., Milly, P. C. D., McMahon, T. A., and Jakeman, A. J.: Effects of rainfall seasonality and soil moisture capacity on mean annual water balance for Australian catchments, Water Resources Research, 41, https://doi.org/10.1029/2004WR003697, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2004WR003697, 2005.

Räisänen, J. and Ruokolainen, L.: Probabilistic forecasts of near-term climate change based on a resampling ensemble technique, Tellus A: Dynamic Meteorology and Oceanography, 58, 461–472, https://doi.org/10.1111/j.1600-0870.2006.00189.x, 2006.

Ratto, M., Castelletti, A., and Pagano, A.: Emulation techniques for the reduction and sensitivity analysis of complex environmental models, Environmental Modelling & Software, 34, 1 – 4, https://doi.org/10.1016/j.envsoft.2011.11.003, 2012.

Ray, D. K., Gerber, J. S., MacDonald, G. K., and West, P. C.: Climate variation explains a third of global crop yield variability, Nature Communications, 6, 5989, https://doi.org/10.1038/ncomms6989, http://www.nature.com/articles/ncomms6989, 2015.

Razavi, S., Tolson, B. A., and Burn, D. H.: Review of surrogate modeling in water resources, Water Resources Research, 48, https://doi.org/10.1029/2011WR011527, 2012.

Riahi, K., Rao, S., Krey, V., Cho, C., Chirkov, V., Fischer, G., Kindermann, G., Nakicenovic, N., and Rafaj, P.: RCP 8.5—A scenario of comparatively high greenhouse gas emissions, Climatic Change, 109, 33, https://doi.org/10.1007/s10584-011-0149-y, 2011.

Roberts, M., Braun, N., R Sinclair, T., B Lobell, D., and Schlenker, W.: Comparing and combining process-based crop models and statistical models with some implications for climate change, Environmental Research Letters, 12, https://doi.org/10.1088/1748-9326/aa7f33., 2017.

Rosenzweig, C., Jones, J., Hatfield, J., Ruane, A., Boote, K., Thorburn, P., Antle, J., Nelson, G., Porter, C., Janssen, S., Asseng, S., Basso, B., Ewert, F., Wallach, D., Baigorria, G., and Winter, J.: The Agricultural Model Intercomparison and Improvement Project (AgMIP): Protocols and pilot studies, Agricultural and Forest Meteorology, 170, 166 – 182, https://doi.org/10.1016/j.agrformet.2012.09.011, 2013.

Rosenzweig, C., Elliott, J., Deryng, D., Ruane, A. C., Müller, C., Arneth, A., Boote, K. J., Folberth, C., Glotter, M., Khabarov, N., Neumann, K., Piontek, F., Pugh, T. A. M., Schmid, E., Stehfest, E., Yang, H., and Jones, J. W.: Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison, Proceedings of the National Academy of Sciences, 111, 3268–3273, https://doi.org/10.1073/pnas.1222463110, 2014.

Ruane, A., I. Hudson, N., Asseng, S., Camarrano, D., Ewert, F., Martre, P., J. Boote, K., Thorburn, P., Aggarwal, P., Angulo, C., Basso, B., Bertuzzi, P., Biernath, C., Brisson, N., Challinor, A., Doltra, J., Gayler, S., Goldberg, R., Grant, R., and Wolf, J.: Multi-wheat-model ensemble responses to interannual climate variability, Environmental Modelling and Software, 81, 86–101, https://doi.org/10.1016/j.envsoft.2016.03.008, 2016.

Ruane, A. C., Cecil, L. D., Horton, R. M., Gordon, R., McCollum, R., Brown, D., Killough, B., Goldberg, R., Greeley, A. P., and Rosenzweig, C.: Climate change impact uncertainties for maize in Panama: Farm information, climate projections, and yield sensitivities, Agricultural and Forest Meteorology, 170, 132 – 145, https://doi.org/10.1016/j.agrformet.2011.10.015, 2013.

Ruane, A. C., McDermid, S., Rosenzweig, C., Baigorria, G. A., Jones, J. W., Romero, C. C., and Cecil, L. D.: Carbon-temperature-water change analysis for peanut production under climate change: A prototype for the AgMIP Coordinated Climate-Crop Modeling Project (C3MP), Glob. Change Biology, 20, 394–407, https://doi.org/10.1111/gcb.12412, 2014.

Ruane, A. C., Goldberg, R., and Chryssanthacopoulos, J.: Climate forcing datasets for agricultural modeling: Merged products for gap-filling and historical climate series estimation, Agric. Forest Meteorol., 200, 233–248, https://doi.org/10.1016/j.agrformet.2014.09.016, 2015.

Ruiz-Ramos, M., Ferrise, R., Rodríguez, A., Lorite, I., Bindi, M., Carter, T., Fronzek, S., Palosuo, T., Pirttioja, N., Baranowski, P., Buis, S., Cammarano, D., Chen, Y., Dumont, B., Ewert, F., Gaiser, T., Hlavinka, P., Hoffmann, H., Höhn, J., Jurecka, F., Kersebaum, K., Krzyszczak, J., Lana, M., Mechiche-Alami, A., Minet, J., Montesino, M., Nendel, C., Porter, J., Ruget, F., Semenov, M., Steinmetz, Z., Stratonovitch, P., Supit, I., Tao, F., Trnka, M., de Wit, A., and Rötter, R.: Adaptation response surfaces for managing wheat under perturbed climate and CO2 in a Mediterranean environment, Agricultural Systems, 159, 260 – 274, https://doi.org/10.1016/j.agsy.2017.01.009, 2018.

Schlenker, W. and Roberts, M. J.: Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change, Proceedings of the National Academy of Sciences, 106, 15 594–15 598, https://doi.org/10.1073/pnas.0906865106, 2009.

Snyder, A., Calvin, K. V., Phillips, M., and Ruane, A. C.: A crop yield change emulator for use in GCAM and similar models:Persephone v1.0, Geoscientific Model Development, 12, 1319–1350, https://doi.org/10.5194/gmd-12-1319-2019, https://www.geosci-model-dev.net/12/1319/2019/, 2019.

Stevanović, M., Popp, A., Lotze-Campen, H., Dietrich, J. P., Müller, C., Bonsch, M., Schmitz, C., Bodirsky, B. L., Humpenöder, F., and Weindl, I.: The impact of high-end climate change on agricultural welfare, Science Advances, 2, https://doi.org/10.1126/sciadv.1501452, https://advances.sciencemag.org/content/2/8/e1501452, 2016.

Storlie, C. B., Swiler, L. P., Helton, J. C., and Sallaberry, C. J.: Implementation and evaluation of nonparametric regression procedures for sensitivity analysis of computationally demanding models, Reliability Engineering & System Safety, 94, 1735 – 1763, https://doi.org/10.1016/j.ress.2009.05.007, 2009.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An Overview of CMIP5 and the Experiment Design, Bulletin of the American Meteorological Society, 93, 485–498, https://doi.org/10.1175/BAMS-D-11-00094.1, 2012.

Tebaldi, C. and Lobell, D. B.: Towards probabilistic projections of climate change impacts on global crop yields, Geophysical Research Letters, 35, https://doi.org/10.1029/2008GL033423, 2008.

Urban, D., Roberts, M. J., Schlenker, W., and Lobell, D. B.: Projected temperature changes indicate significant increase in interannual variability of U.S. maize yields: A Letter, Climatic Change, 112, 525–533, https://doi.org/10.1007/s10584-012-0428-2, 2012.

von Bloh, W., Schaphoff, S., Müller, C., Rolinski, S., Waha, K., and Zaehle, S.: Implementing the Nitrogen cycle into the dynamic global vegetation, hydrology and crop growth model LPJmL (version 5.0), Geoscientific Model Development, 11, 2789–2812, https://doi.org/10.5194/gmd-11-2789-2018, 2018.

Waha, K., van Bussel, L. G. J., Müller, C., and Bondeau, A.: Climate-driven simulation of global crop sowing dates, Global Ecology and Biogeography, 21, 247–259, https://doi.org/10.1111/j.1466-8238.2011.00678.x, https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1466-8238.2011.00678.x, 2012.

Warszawski, L., Frieler, K., Huber, V., Piontek, F., Serdeczny, O., and Schewe, J.: The Inter-Sectoral Impact Model Intercomparison Project (ISI–MIP): Project framework, Proceedings of the National Academy of Sciences, 111, 3228–3232, https://doi.org/10.1073/pnas.1312330110, 2014.

Weedon, G. P., Balsamo, G., Bellouin, N., Gomes, S., Best, M. J., and Viterbo, P.: The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data, Water Resources Research, 50, 7505–7514, 2014.

Wiebe, K., Lotze-Campen, H., Sands, R., Tabeau, A., van der Mensbrugghe, D., Biewald, A., Bodirsky, B., Islam, S., Kavallari, A., Mason-D'Croz, D., Müller, C., Popp, A., Robertson, R., Robinson, S., van Meijl, H., and Willenbockel, D.: Climate change impacts on agriculture in 2050 under a range of plausible socioeconomic and emissions scenarios, Environmental Research Letters, 10, 085 010, https://doi.org/10.1088/1748-9326/10/8/085010, https://doi.org/10.1088%2F1748-9326%2F10%2F8%2F085010, 2015.

Williams, K., Gornall, J., Harper, A., Wiltshire, A., Hemming, D., Quaife, T., Arkebauer, T., and Scoby, D.: Evaluation of JULES-crop performance against site observations of irrigated maize from Mead, Nebraska, Geoscientific Model Development, 10, 1291–1320, https://doi.org/10.5194/gmd-10-1291-2017, 2017.

Williams, K. E. and Falloon, P. D.: Sources of interannual yield variability in JULES-crop and implications for forcing with seasonal weather forecasts, Geoscientific Model Development, 8, 3987–3997, https://doi.org/10.5194/gmd-8-3987-2015, 2015.

Zabel, F., Delzeit, R., Schneider, J. M., Seppelt, R., Mauser, W., and Vàclavík, T.: Global impacts of future cropland expansion and intensification on agricultural markets and biodiversity, Nature Communications, 10, 2844, https://doi.org/10.1038/s41467-019-10775-z, http://www.nature.com/articles/s41467-019-10775-z, 2019.

Zhao, C., Piao, S., Wang, X., Huang, Y., Ciais, P., Elliott, J., Huang, M., Janssens, I. A., Li, T., Lian, X., et al.: Plausible rice yield losses under future climate warming, Nature plants, 3, 1–5, 2016.

Zhao, C., Liu, B., Piao, S., Wang, X., Lobell, D. B., Huang, Y., Huang, M., Yao, Y., Bassu, S., Ciais, P., Durand, J. L., Elliott, J., Ewert, F., Janssens, I. A., Li, T., Lin, E., Liu, Q., Martre, P., Müller, C., Peng, S., Peñuelas, J., Ruane, A. C., Wallach, D., Wang, T., Wu, D., Liu, Z., Zhu, Y., Zhu, Z., and Asseng, S.: Temperature increase reduces global yields of major crops in four independent estimates, Proc. Natl. Acad. Sci., 114, 9326–9331, https://doi.org/10.1073/pnas.1701762114, 2017.