Dear Dr. Sweet and co-authors,

Thank you for submitting your revised manuscript and the point-by-point responses to the reviewers' comments. As neither reviewer has been available to provide a second-round review of your revision, I am making the editorial decision based on my own assessment of the revised work in light of the original reviews.

The manuscript introduces a novel and well-motivated two-stage framework for identifying a small number of parsimonious, interpretable climate drivers of agricultural yield failure from daily meteorological data. The use of two independent process-based crop models (pDSSAT and LPJmL) as a validation testbed, together with the application to observed US county-level maize yields, provides a rigorous and comprehensive evaluation. The paper makes a genuine contribution to climate impact science and to the methodological literature on data-driven feature identification.

The revision has substantially addressed the principal concerns raised by both reviewers. In particular: Reviewer 1's central concern about the spatial transferability of the identified drivers has been convincingly resolved. Reviewer 2's requests for clearer specification of candidate pools and aggregation functions, clarification of the odds ratio interpretation, and justification of the PRISM and ISIMIP variable selections have all been addressed.
However, there are still a few aspects that are not fully addressed. Before I can accept the manuscript for publication in GMD, I request that the authors address the following points in a minor revision.

Thank you for taking the time to assess our revised manuscript. We are glad that you feel our prior revisions have addressed the principal concerns of the reviewers, and we really appreciate the constructive feedback. We are in full agreement with the suggestions made, and have revised the manuscript accordingly. Responses to each specific suggestion, and descriptions of the relevant changes made to the manuscript, lie in more detail below.

Suggested changes:
1. Integration of the three-test-set performance results into the manuscript. The table summarizing model performance across the temporal, spatial, and spatiotemporal test sets, as presented in the author response to Reviewer 1, constitutes a core result of the revised paper and should appear in the main manuscript — either as a formal table after some modification in Section 4.2 or as an appropriately referenced supplementary table. These numbers will be helpful for readers to understand the spatial and temporal transferability of the identified drivers.

We have added an adapted version of this table **(Table 1)** to the main manuscript which also includes the scores for LPJmL. Correspondingly, in the main body of the text, we

have removed or altered the sentences reporting these individual scores to improve readability, and added a summarising sentence.

Revised text (lines 256-267):

> A logistic regression model fit on the training data using ten identified climate drivers of pDSSAT yield failure as predictors is able to achieve a ROC AUC of 0.83, average precision of 0.32, Brier score of 0.056 and log loss of 0.20 on the held-out test years. **Comparable scores are achieved when applied to the held-out spatial test set (covering the same growing seasons as used for identifying drivers and training the model) and the spatiotemporal test set (including only held-out years and locations) (Table 1). Notably, the model is able to outperform lasso logistic regression models using all baseline climate feature sets tested for all metrics calculated and for all held-out sets.** The model captures the year-to-year variability in yield failure well (Fig. 4a), with a Pearson correlation of 0.84 between the true and predicted annual proportions of grid cells experiencing yield failure over the training region (above 35 degrees latitude) and 0.91 over the held-out region (below 35 degrees latitude). Notably, model performance does not appear to deteriorate over time: annual ROC AUC scores slightly increase over the decades, while Brier score and log loss decrease. Maps of predictions and ground truth in 2011, in which the majority of grid cells in the lower latitudes experienced a yield failure (Fig. 4c,d) show general agreement in spatial variability, including over the held-out lower latitudes in which yields were most adversely impacted.

> *2. Reporting of the threshold sensitivity analysis. The authors state in their response that sensitivity analyses were performed using 5% and 20% (5th & 20th percentiles) yield failure thresholds, with no large differences observed. This finding should be stated explicitly in the manuscript (e.g., with a sentence in Section 3.1 or Section 4.3), ideally with a brief quantitative summary and a reference to a supplementary figure/table. As stated, the choice of the 10th percentile remains theoretically unmotivated in the text, and even a brief empirical justification would strengthen the paper's methodological transparency.*

We agree that including threshold sensitivity analysis results would strengthen the paper's methodological transparency. We also think it provides further evidence that our method is robust to such choices. Rather than report the results from our previous sensitivity tests at 5% and 20%, we have now executed a much more thorough analysis in which we reproduce all of the main results from the paper using 5th and 15th percentile thresholds, while keeping all other parameters (method parameters, models, metrics etc) identical to those used when obtaining the main results. We have included figures summarising these results in a new appendix (**Figs C1-4**) and referenced these in the text.

For all thresholds, the results confirm that our method is able to consistently identify plausible climate drivers that can outperform baseline sets of climate indices on held-out years and/or regions, and this is robust to performance metric, model type, crop model, and whether or not detrending is applied.

We have added a sentence describing this to the Methods section, and some paragraphs discussing the results from this analysis in more detail in the Results and Discussion sections.

Revised text (lines 124-140):

> Each datapoint in our compiled dataset consists of a binary target variable and the corresponding daily multivariate climate input for one growing season at one grid cell. Due to the trends in yields due to improved management practices and breeding that have occurred over the last decades, researchers often consider yield anomalies (after detrending) rather than their absolute values. We test our method on both non-detrended and detrended yields (based on subtracting a seven-year rolling mean, meaning that we discard datapoints from the first six years), in order to assess the utility of the method for analysis of both relative yield failure and the occurrence of yields below an absolute threshold. The target variable is maize yield failure, which we define as any yield below the 10th percentile observed during the training period at that grid cell. **We assess the robustness of our method to this choice by reproducing the results using 5th and 15th percentile thresholds.** We test our method on both non-detrended and detrended yields (based on subtracting a seven-year rolling mean, meaning that we discard datapoints from the first six years). The dataset is split into a training set (the first 30\% of growing seasons, covering 1902-1937 for non-detrended yields and 1907-1940 for detrended, and all grid cells with latitudes above 35 degrees north) and three test sets: a temporal test set (the last 70% of growing seasons, covering 1938-2016 and 1941-2016, for grid cells of latitude above 35 degrees), a spatial test set (the first 30% of growing seasons in grid cells below 35 degrees) and a spatiotemporal test set (the last 70% of growing seasons in grid cells below 35 degrees). Note that, for non-detrended LPJmL yields, the 10th percentile over the training years was zero for 167 grid cells. These low-yielding areas are removed from the training and test set. After processing, the training sets are made up of 50,000 to 60,000 datapoints, the temporal test sets consist of 115,000 to 135,000, the spatial test sets 13,000 to 18,000, and the spatiotemporal test sets 30,000 to 39,000.

Additional text (lines 357-376):

> The results described so far are based on applying the method to identify drivers of yield failure, where yield failure is defined as yields below the 10th percentile of those occurring during the training years. When reproducing the results using a more extreme (5th percentile) or more moderate (15th percentile) definition of yield failure, we find that the variables, aggregations and time-durations of identified drivers are largely identical, with some minor differences. In general, the

results suggest that the method is slightly better able to eliminate variables that were not used as input for the crop models when using higher (more moderate) thresholds for defining yield failure. For example, when using a 5th percentile threshold, specific humidity is sometimes identified as a driver for LPJmL, but not when the 10th or 15th percentile is used.

The ten identified drivers based on 100 repeats are largely similar for all yield failure thresholds tested (Figs. C1-C2). In particular, the precipitation drivers and those using the number of days where maximum temperature is above the 90th percentile are almost identical. Almost all drivers identified are based on precipitation, minimum or maximum daily temperatures at different stages of the growing season. However, there are some differences, particularly in the drivers identified based on the 5th percentile, which features long-wave radiation for pDSSAT and specific humidity for LPJmL.

If a more extreme definition of yield failure is used (5th percentile), we observe generally poorer predictive performance across the held-out test sets, particularly in terms of average precision, for all models, whether using identified drivers or baseline climate feature sets (Fig. C3). This is likely because the task and the training set becomes more imbalanced, by definition, with fewer examples of yield failure. Correspondingly, when using a (more moderate) 15th percentile threshold, we observe overall higher scores across the models, metrics and test sets (Fig. C4). When comparing the relative performance of models using our drivers to those using baseline feature sets, however, the results are essentially unchanged for all thresholds: models using drivers identified using our proposed method as predictive variables usually outperform those using standard sets of climate indices.

Revised text (lines 436-454):

Many variables which influence agricultural yield are correlated in space and among themselves, which can lead to errors in empirically-estimated relationships. Climate change causes warmer temperatures and shifts in precipitation patterns, strongly affecting agricultural regions (Lobell and Di Tommaso, 2025). Data-driven models have difficulties extrapolating outside of the observed distribution, which could impede their use for projecting impacts under future climate change scenarios. The robustness of our results, despite the challenging split between training and test years and regions used in this study, show that our approach is successful in tackling these challenges. Data-driven emulators of process-based crop models have been proposed to efficiently downscale simulations and facilitate the creation of large ensembles of agricultural projections in different climate scenarios. Published emulators of crop models have used predictors such as the mean and standard deviation of temperature and precipitation over the growing season or per month, and measures of climate extremes such as the number of extreme degree days or the maximum consecutive days with no rainfall (Folberth et al., 2019; Liu et al., 2023). However, due to the presence of spatiotemporal autocorrelation in the

data used for training these models, emulator skill has been found to drop sharply when tested on held-out years or locations (Liu et al., 2023; Sweet et al., 2023). Our approach, on the other hand, identifies a set of climate features that are robust drivers of a specific impact or metric and can be used to create a lightweight, interpretable model that achieves high performance on datapoints many decades after the training period or in unseen latitudes. Such emulators or metamodels could also be used to analyse the processes leading to specific model outcomes which might emerge from the interaction of multiple complex mechanisms and therefore not be well-understood. **Our results also show that our method is robust to the percentile threshold used for defining yield failure. This suggests that the method could be used by researchers to differentiate between drivers of more extreme or more moderate climate impacts, which could be caused by different processes.**

We also include Jupyter notebooks that reproduce all the relevant manuscript figures using the different thresholds (identified drivers using both detrended and non-detrended data, performance of models using identified drivers versus baseline sets, spatiotemporal performance of logistic regression models using the identified drivers over the held-out years and locations) in the published code repository for readers that are interested in a detailed comparison.

3. Clearer treatment of the detrended pDSSAT discrepancy in the manuscript. The response to Reviewer 2 acknowledges that shortwave radiation appears as a driver when yields are detrended for pDSSAT, whereas it is eliminated in the non-detrended case. The authors explain that it is the last of the ten drivers to be selected, but acknowledge that, "the point made by the reviewer is an important one that we should, and will now, discuss further in the manuscript." I didn't find an explicit discussion of this in the manuscript. Please include some discussion to help readers understand what this sensitivity to detrending implies for the interpretation of the identified drivers and the robustness of the method more generally.

Thank you for raising this issue. After exploring this discrepancy further, in combination with the results from our analysis of the results using different yield failure thresholds, we think that this effect is due to the added difficulty of the task when detrending the impact data. We now include some more extensive discussion of this complicated issue in the Discussion and Conclusions section.

Additional text (lines 486-506):

While our results suggest that this method can generally identify robust and plausible drivers, its performance (like any data-driven approach) depends on task difficulty and data availability. This is illustrated when varying the threshold used to define yield failure. A more extreme definition (i.e, a 5th percentile threshold) means that the training set will contain fewer examples of yield failure,

which could make it harder to identify robust drivers. This may explain why, for LPJmL, specific humidity (which is not an input to the crop model) is sometimes identified as a driver when using a 5th percentile threshold, but not for less extreme thresholds. Therefore, drivers identified at very extreme impact definitions should be interpreted with caution. One way to assess whether the identified drivers reflect genuinely different underlying processes at those levels is to compare their predictive performance against drivers identified using more moderate thresholds. If the drivers identified for more extreme events are spurious due to limited and imbalanced data, then drivers identified under easier conditions would likely perform better even when predicting more extreme events.

Detrending impact data introduces another layer of difficulty, as identical climate conditions can lead to yield failures in one year or location, but not in another, depending on yields experienced in previous years (Siebert et al. 2017). Consistent with this, our results show generally degraded performance scores for all models (whether using identified drivers or baseline climate feature sets) for detrended data, in line with this. In fact, even a perfect emulator of the underlying crop model could not achieve perfect predictive performance without access to information in previous years. As a result, spurious correlations may be more likely to achieve predictive skill and robustness comparable to that of true causal drivers. This could explain why specific humidity appears as an identified driver for detrended pDSSAT, but not in the non-detrended case. In line with this, we also find that as the yield failure threshold becomes less extreme (meaning the task becomes easier and the training set includes more examples of yield failure), specific humidity appears less frequently in the identified drivers. It is important to note that this reduction is much stronger for the identified drivers than for the features selected in the first step of the method, suggesting that the full method is still more reliable than sequential feature selection alone.

4. The justification provided for n=100 bootstrap repeats — that "initial testing indicated that 100 repeats sufficiently approximate larger sample sizes" — would be strengthened by a brief quantitative statement (e.g., the threshold n at which the results stabilized, or a reference to a convergence test in the appendix). It would not suffice to just say "practicality reasons". Nowadays, there is barely any computational restriction why not to use a larger bootstrap such as n=1000 or 10000.

We have added a figure **(Figure A1)** in the appendix which plots, for n=10 to 1000, the estimated 25th, 50th and 75th percentiles of lasso logistic regression model performance scores as measured using the four different metrics on the temporal test set (using 5, 10 or 15 identified drivers) for LPJmL. These show consistent stability from n=100 for all metrics tested. We observed similar results for pDSSAT.

Revised text (lines 192-199):
We further test the robustness of the approach by conducting 100 bootstrap repeats, sampling 20 pools (with replacement) from the 100 pools described above. For each repeat, we extract sets of ten climate drivers from the 20 pools, and evaluate the extent to which these sets consist of the variables actually taken as input by the crop models by counting the variables used by the features collected from each pool (in Step 2 of the driver identification method), and from the condensed drivers over all bootstrap repeats. Second, for each of the repeats, we construct models with 5, 10 or 15 identified drivers and evaluate their predictive performance over the three test sets, in comparison to a variety of commonly-used baseline feature sets. **An assessment of model performance estimates using up to 1,000 bootstrap repeats, for Lasso logistic regression models on the temporal test set for LPJmL, suggests that 100 repetitions is sufficient for stable model performance estimation (Fig. A1).**

> *Minor suggestions (optional but recommended):*
> *5. More recent energy-based PET formulations. The authors' response to Reviewer 2 correctly notes that both pDSSAT and LPJmL use the Priestley-Taylor approach to compute potential evapotranspiration, which avoids the need for explicit humidity or windspeed inputs and thereby influences which variables appear as climate drivers. While this justification is sound, readers should be aware that the Priestley-Taylor equation is a relatively simple, decades-old approximation and that more recent energy-based PET formulations have been proposed that refine this framework while similarly requiring neither humidity nor wind data (e.g., Yang & Roderick, 2019, doi:10.1002/qj.3481; Milly & Dunne, 2016). There are even better energy-based PETs that incorporates humidity (Zhou & Yu, 2024, Journal of Hydrology; Kim, Y., et al. 2023. Earth's Future). The authors may wish to note briefly in their Discussion that future crop model implementations using more physically complete PET schemes could alter the set of variables identified as drivers, and that this is a further motivation for applying the method to outputs from a diverse range of process-based models.*

Thank you for raising this. This is a useful, specific example of a crop model component that the method could be evaluated on in future work, and we agree it is good motivation for applying the method to simulations from diverse models. We have now included some sentences on this.

Revised text (lines 464-485):
> However, these interpretations should be considered with caution. Both of the crop models used consist of multiple mechanisms that can interact in surprising ways, and the process by which we transform yields to yield failures is dependent on the grid cell and time period selected. This means that, even in this simulated setup, confirming that climate drivers identified by our approach are `accurate' is not straightforward. Furthermore, while we evaluate our method based on the principle that identified drivers should consider only the climate variables taken as input by each crop model, this does not take into account the fact that secondary variables not used as input may be internally estimated by the crop model using

other information. For example, both crop models estimate potential evapotranspiration using the Priestley-Taylor radiation-based approach, which does not consider air humidity but implicitly represents atmospheric demand through available energy and an empirical coefficient. The match between selected and input variables is only one aspect of the method's fidelity, and should be considered in combination with the sensitivity of crop model processes used to compute yield to individual driver variables, and consistency of the identified drivers with the known underlying processes in terms of phenological timing and the aggregation strategy employed, as well as the climate variable used. However, these criteria are difficult to quantify, and require good understanding of the underlying crop model functionality to evaluate. **Other formulations for potential evapotranspiration are employed in crop models (e.g. Cammarano et al. 2016) and applying the method to simulations from a more diverse set of crop models in general would likely find other variables as important.** Out-of-distribution model performance can also provide evidence that drivers obtained are a good approximation for crop model behaviour.

6. The updated Figure 4 now displays both the training-region time series (panel a) and the held-out lower-latitude time series (panel b), which is a significant addition. Please verify that both the figure caption and legend unambiguously identify which panel corresponds to which region and test set (better add a text label in each panel), and which shaded region corresponds to which training period (e.g., currently both a and b have shaded regions but only a is explained in caption; better clearly state the training period in caption).

Thank you for noticing these issues and for the helpful suggestions. We have updated the relevant figures with text labels denoting the training set and the temporal, spatial and spatiotemporal test sets (Figures 4, A3, B3, B4) and we hope they are now easier for readers to understand. We have also updated the caption for Figure 4 to be more explicit.

Revised caption (Fig. 4):

Predictions of a logistic regression model using ten identified pDSSAT drivers **on held-out years and regions**. (a) True (black) and predicted (orange) proportion of grid cells experiencing yield failure each year over the grid cells used for training (latitudes above 35 degrees), using a threshold of 0.235 to define a yield failure prediction (selected based on achieving the maximal f-score of 0.405 over the training years, which are marked by the grey shaded area). (b) True (black) and predicted (orange) proportion of grid cells experiencing yield failure each year in the held-out grid cells (latitude below 35 degrees). **Unshaded years represent the spatial test set (held-out regions during years included in training data) and shaded years represent the spatiotemporal set (held-out regions and years).** Model predictions for 2011 are illustrated in panel (c), with

orange dots marking grid cells where the prediction exceeds the threshold and therefore yield failure is predicted. The corresponding ground truth is shown in panel (d), with black areas indicating locations experiencing yield failure. **The curved line indicates the spatial split used to define the test sets. No data from locations below the line, from any year, is used for identifying drivers or training the model.**