“Flood damage functions for rice: Synthesizing evidence and building data-driven models”

Response to the Reviewers

Reviewer 1:

We very much appreciate the time and effort that the reviewer dedicated to providing feedback on our manuscript and are grateful for the comments and suggestions for improvements. We have incorporated the suggestions. Below, we provide point-by-point responses to the reviewer’s comments.

Comment #1: The study models rice yield loss using different approaches. It uses generalized and localized splits of the data. It probes deterministic, probabilistic, Bayesian regression, and random forest regression types of models. Based on my evaluation, I have serious doubts about the claim that the authors made regarding the results. When you test models making different splits, it is wrong to choose the best-performing model as the best. ***You need to focus on the models that have been evaluated to reduce bias through independent evaluation****. In this case,* ***the authors chose the generalized model as the best, but in my opinion, their selection is biased because it uses all the data****. Instead, they should focus on the localized split (CRV). The models called "transferred" are the ones that give an independent evaluation.*

Response #1.1: We thank the reviewer for this important comment. All models were evaluated using splits that kept training and testing observations independent. We compared generalized models only against other generalized models, and localized models only against other localized models, based on average performance over many independent splits. We do not claim that generalized models outperform localized models overall. The localized transfer evaluations (CRV) were included to show performance on unseen regional conditions (when models are moved across regions). All models are provided openly so end-users can choose the appropriate option for their application (see Data Availability Statement). We have added a graphical abstract that visualizes the validation steps and revised Section 2.3.

Change made to the manuscript:

2.3 Model validation

We evaluated model performance following the four model validation steps shown in Fig. 3. First, we conducted 10-fold cross-validation (CV) (Fig. S3) to assess predictive accuracy. For the 10-fold CV, we randomly split the observed rice yield loss data into ten folds of roughly the same size. Each fold served as a validation set for a model that was fitted with the remaining observations in the training set (James et al., 2013). Second, we tested model transferability across locations, also referred to as cross-region validation (CRV), by training the models on damage data from region A and validating them on data from region B. Third, we tested the performance of the generalized models in each region as part of the 10-fold CV. In this step, we used all observations per left-out fold from one region for the validation and calculated mean performance metrics across the 10 folds. Fourth, the models trained on the Shrestha et al. data from Myanmar were compared with the ramp functions by Shrestha et al. (2021). After the validation, leave-nothing-out (LNO) models were trained on the full dataset for future application. Table 4 summarizes the calibration and validation strategies for each model type.

Performance was assessed using three established metrics: mean absolute error (MAE), mean bias error (MBE), and continuous ranked probability score (CRPS). These allow for comparison across deterministic and probabilistic models and align with prior flood damage modeling studies (e.g., Schoppa et al., 2020; Gneiting and Katzfuss, 2014). Full definitions and formulas of the performance metrics are provided in the Supplementary Information (Section 3). By examining how performance metrics vary across folds in the 10-fold CV, we gained insights into the stability and robustness of the models. Low variability (tightly clustered errors) indicates model stability, while high variability (widely spread errors) indicates model instability, meaning that the model is not generalizing well.

Response #1.2: The generalized models were evaluated using 10-fold cross-validation (first row in Table 4). When providing models for future applications, we then retrained the generalized models on the full dataset using the Leave Nothing Out (LNO) approach. The LNO models were not used for validation and do not contribute to any of the reported performance metrics. We have revised Table 4 to make this distinction explicit.

Changes made to the manuscript (in three sections):

3.2.1 Performance of the generalized models in a 10-fold cross-validation (CV)

Among the generalized models, RF achieved the best performance in a 10-fold CV conducted with a joint dataset combining observations from Myanmar and Thailand:

4. Conclusion:

Our findings highlight that, among the generalized models, the Random Forest (RF) model performs best, followed by the multivariate Bayesian Regression Model (BRM).

Data availability:

The generalized models and localized models for Thailand and Myanmar built in this paper are available on GitHub ...

Comment #2: I also recommend to the authors to evaluate the models spatially and to use, at least, the locations to investigate how dependent on the locations the models are. This could supply insights into other variables that could be used in the future to improve the models.

Response #2: We thank the reviewer for the comment. We evaluated the models spatially by assessing how well each model performs when transferred across regions (trained in one region and tested in another region). We also assessed the performance of the generalized models locally (by calculating the performance metrics in the 10-fold CV of the generalized model separately for testing data from each region). The results of these assessments are shown in Fig. 5 (points and triangles). We added an explanation in Section 2.3. We did not include region as a predictor variable because this would constrain the generalized models to only a very small set of locations and make them mathematically equivalent to the localized models that already treat damage by region category. Additionally, precise spatial attributes (e.g., latitude-longitude coordinates) were not available in the source survey data.

Changes made to the manuscript (Section 2.3): See Response #1.1.

Comment #3: Majors.

Methodology

The methodology is briefly explained. It should be extended for a better comprehension of the reader.

Regarding model evaluation, from my perspective, the only useful evaluations are the LOOCV and the CRV. The first lets you assess model stability by analyzing the variability of the 10 folds you receive. If it is shrunk, it is stable; if it is sparse, it is unstable. These results should be discussed. In the second evaluation, you provide an independent and useful analysis by using data from other locations and validating it against unknown locations. This is the real truth of how the model works.

Response #3: We thank the reviewer for the suggestions concerning the methodology. The only models used for evaluation are the 10-fold CV and the CRV. We revised Table 4 for clarification to highlight that no validation was conducted for the retrained generalized models trained with all the data (LNO), as these only serve for future applications. We have expanded Section 2.3, including details previously in the Supplementary Information.

Changes made to the manuscript (Section 2.3): See Response #1.1.

Comment #4: Regarding the CRV, you can also use ten folds; however, I do not understand why only one dot is shown in Fig. 3. The results presented in this figure, in my opinion, do not support your conclusions. The best-performing models, based on the CRPS, are BRM of SDF (prob). You cannot assign equal weight to the distribution of the ten folds for just one point evaluation (it is probably biased). You should run a ten-fold evaluation for all the models and compare them fairly.

Response #4.1: We thank the reviewer for the comment and have made several revisions.

We performed **10-fold CV** for both the generalized model (trained with data from both regions) and the localized models (each trained with data from one region) (boxplots in the previous Fig. 3, now Fig. 5). In the CRV ***for the localized models***, we trained the model on all observations from region A and evaluated it on all observations from region B, holding the test region (region B) completely unseen. Since we have only two regions available, this results in two distinct CRV runs (A→B and B→A). Therefore, each model yields a single point per CRV evaluation. We did not implement 10-fold CV for CRV, given that with only two regions, applying 10-fold cross-validation is not feasible in a traditional sense (i.e., each fold is one region). Furthermore, since we are using global metrics such as CRPS, MSE, and MAE - which are additive and stable over the full test set - evaluating on the entire test region is mathematically equivalent to averaging across smaller test subsets.

We agree it is important to assess variability in performance. Therefore, in addition to the CRV analysis, we conducted 10-fold cross-validation using the generalized model trained on data from both A and B to test the model performance per region (step 3 in the validation steps presented in Fig. 3). During this analysis, we computed and reported CV errors separately for data originating from A and from B. This allowed us to quantify region-specific generalization under shared training. The results have been updated in the revised Fig. 5, where the shapes represent mean performance across 10 folds, and the thin boxplots show the distribution of fold-level performance metrics for Thailand ("TH") and Myanmar ("MM"). In addition, we expanded Section 3.4 to highlight an initial observation from the fold-level variability shown in Fig. 5.

Changes made to the manuscript (Section 3.4 & Section 3.2.1):

Section 3.4: The spread of the MAEs of the generalized BRM across ten folds is smaller in the model validation in Myanmar than in Thailand, indicating less stable results in Thailand (thin boxplots in the "generalized models" column in Fig. 5).

Section 3.2.1: Fig. 5: Results of performance evaluation and transferability assessment. Results are shown for three performance metrics (rows) and four calibration setups (generalized, localized trained on Thailand data and Myanmar data, and localized trained on Myanmar data by Shrestha et al. 2021) (columns). Colors indicate the model types (RF, BRM, probabilistic SDF, and deterministic SDF). Boxplots show the 10-fold cross-validation results. The thick lines show the median; boxes represent the interquartile range (IQR); whiskers extend to 1.5×IQR; and black points are outliers. Each box summarizes variability across 10 folds. For localized models, shapes indicate the performance of transferred models (CRV). For generalized models, shapes indicate mean performance metrics from local validation across ten folds, and the thin boxplots summarize variability in the local validation across the 10 folds. The lines mark the mean performance of the benchmark ramp function in Myanmar (solid red line) and in Thailand (dashed blue lines).

Response #4.2: We also clarified that "best performance" of the RF model refers only to being the best performing model under criterion 1, and only in comparison with the other four generalized models evaluated using the same 10 fold unseen test splits.

Changes made to the manuscript (Section 3.2, Section 3.2.1, and Section 4):

3.2 Model performance

For each metric, this study examined model performance based on two criteria: firstly, based on local data (using testing data from the same region, where the training data originated from) in a 10-fold CV (boxplots in Fig. 5), and, secondly, in a CRV (shapes in the localized model columns of Fig. 5; Section 3.4). The performance of generalized models in a specific region was assessed by separately calculating the performance for the observations per region in the left-out folds of the 10-fold CV (shapes and thin boxplots in the left column of Fig. 5). A detailed summary of the model performance assessment is provided in Table S6. This subsection presents the performance of the generalized models based on the first criterion (3.2.1) and compares their performance with the ramp functions (3.2.2).

3.2.1 Performance of the generalized models in a 10-fold cross-validation

Among the generalized models, RF achieved the best performance in a 10-fold CV conducted with a joint dataset combining observations from Myanmar and Thailand:

- MAE: RF had the lowest error (20.3%), followed by BRM (22.3%), probabilistic SDF (24.4%), and deterministic SDF (26.3%).
- MBE: All models showed low bias, with RF (0.1%) having the lowest bias, followed by BRM (0.2%), deterministic SDF (0.4%), and the probabilistic SDF (0.5%).
- CRPS: RF also achieved the lowest CRPS (12.4%), indicating stronger probabilistic prediction performance than the BRM (14.2%) and the probabilistic SDF (15.4%).

In the 10-fold CV for generalized models, model complexity improved performance consistently: multivariable models (RF and BRM) outperformed the simpler univariable SDFs.

4. Conclusion

Our findings highlight that, among the generalized models, the Random Forest (RF) model performs best, followed by the multivariate Bayesian Regression Model (BRM).

Comment #5: Generalized models perform better because of overfitting. Again, the independent evaluation (CRV) tells you how the model could perform in unknown conditions.

Response #5: We very much appreciate the comment and updated Sections “2.3 Model validation” and “3.2 Model performance” to clarify that two different criteria were used for the performance evaluation. We are not claiming that the generalized models perform better than local models, but rather we assess and compare the performance only between different generalized models (e.g., RF vs. BRM) and only between localized models (applied locally and transferred). All models (localized & generalized models) were made available on GitHub (see link in the data availability statement).

Comment #6: Section 3.5.1 seems irrelevant to this study. It is not clear how it supports the results obtained.

Response #6: We thank the reviewer for this comment. Section 3.5.1 presents an inventory of existing flood damage models for rice, introduced to clearly ground the model types we test and to explicitly identify current research gaps. Based on this feedback, this section has now been moved to the introduction (Section 1.3) and reframed to focus on:

1. Lack of probabilistic flood damage models for rice. All models found in the literature are deterministic.
2. Limited evidence on validation in prior work. Only 40% of studies present model validation results.
3. Lack of evidence on model transferability.

The inventory is no longer included in the Results section, and therefore does not serve as a performance claim itself, but as context motivating our work.

Changes made to the manuscript (Section 1.3):

Model validation was reported for 40% of all models, and model transferability was only tested for a single model, the ramp functions by Shrestha et al. (2021). This study addresses these gaps by conducting model validation and transferability assessments. Less than half of the models incorporate growth stage as a predictor. Two-thirds of the models incorporate flood duration as a predictor, primarily as a categorized variable (60% of all models) and rarely as a continuous variable (15% of all models). The developed models in this study use flood duration as a continuous input variable. No model offers probabilistic outputs or formal uncertainty analysis. This highlights the need for more data-driven, multivariable, and transferable models. Most flood damage models for rice define the response variable in relative terms. In line with the existing literature, the models developed in this paper predict relative yield loss.

Comment #7: Conclusion

The conclusion is excessively detailed; it should provide a concise summary of the key findings. The first paragraph corresponds to the introduction. I have serious doubts about the conclusion regarding which figure the authors are referring to (40,000 ha). extract from the results. It is not strongly supported by the results shown in the manuscript.

Response #7: We thank the reviewer for this feedback. We streamlined the conclusion and removed some of the context that was repeated from the introduction. We have further reviewed the conclusion in light of the changes and clarifications made in response to previous comments, and we have adapted it as needed to better align with results and discussion.

Comment #8: Minors

L145: A comma is used to separate thousands. Please clarify which figure you are referring to when mentioning 40,00 ha.

Response #8: We appreciate the comment. 40,000 hectares of paddy fields were affected by flood events in the Bago River Basin in Myanmar, according to Shrestha et al. 2021. Paddy fields account for about half of the sown area in the basin. We decided to focus on the share of paddy fields of the sown area and removed the reference by Shrestha et al. 2021.

Changes made to the manuscript (Section 2.1):

In the Bago River Basin in Southern Myanmar, paddy fields account for about half of the sown area. The 331 km long Bago River is used for hydropower generation, irrigation, and fishing (Win et al., 2018).

Comment #9: Table 3: What is the value of h saturation? The table only displays the value of h min.

Response #9: We thank the reviewer for the suggestion to add the value of h saturation. We added the value of $h_{saturation}$ = 314 cm for the linear model and the value of $h_{saturation}$ = 140 cm for the univariable Bayesian regression model (det-SDF) to Table 3.

Comment #10: Table 3: Could you please clarify where exactly the supplementary information is?

Response #10: We revised Table 3, which now specifies the Sections in the Supplementary Information in the right column.

Comment #11: Fig. S1 should be in the manuscript. Fig S2, should be in the manuscript. It allows the reader to have a clear view of the variables used in the models.

Response #11: Thank you for your suggestion to move the contextual map from the Supplementary Information into the main text. While we agree that the map is informative and provides useful background, we consider it not essential to the core findings or analyses of the paper. The main manuscript already includes six figures, several of which are multi-panel and data-dense, and adding another figure may compromise the clarity and flow of the results. To ensure accessibility, we have ensured the map remains prominently available in the Supplementary Information and have referenced it clearly in the main text. We hope this strikes a good balance between providing helpful context and maintaining focus in the main narrative.

Change made to the manuscript (Section 2.1):

We interviewed 584 households (20% of the 2,904 total) in the Lower Songkhram River Basin in March 2023, exceeding the minimum sample size for a 95% confidence level (see a map in Fig. S1 and Tables S1-S2 for details on the data collection).

Text referring to Fig. S2 in the manuscript (Section 2.2):

The considered phenological traits (duration of each plant growth stage and the plant height at each plant growth stage) are similar in the datasets used to develop the models (see Fig. S2).

Comment #12: Table 3: The models should be briefly described in the manuscript rather than just shown in the supplementary.

Response #12: We thank the reviewer for the suggestion. We have added an explanation for each model in the text (Section 2.2) and summarized key details for each model in Table 3.

Comment #13: L173-174: What is the purpose of running a model with LNO?

Response #13: The LNO models were not used for validation or for reporting model performance. They were retrained on the full dataset only after 10-fold model validation to provide models for future application. We revised Table 4 to make this explicit.

Comment #14: Fig. 4. These are boxplots, not violin plots.

Response #14: We thank the reviewer for spotting this error and have corrected the title of Fig. 6 (previously Fig. 4).

Change made to the manuscript (Section 3.3):

Fig. 6: Box plots indicating the predictor importance for the generalized and localized RF models. Each boxplot shows the median (thick horizontal line), the interquartile range (IQR), the most extreme values within 1.5xIQR (whiskers), and outliers (black points).

Comment #15: Fig. 6 is described before Fig. 5.

Response #15: We have moved the figure showing the inventory of flood damage models for rice to the introduction and have updated the figure numbering accordingly.

Reviewer 2:

We very much appreciate the time and effort that the reviewer dedicated to providing feedback on our manuscript and are grateful for the comments and suggestions for improvements. We have incorporated the suggestions. Below, we provide point-by-point responses to the reviewer's comments.

Comment #1: This study addresses the issue regarding the impact of floods on rice crops and presents the contribution through model evaluation. However, there are some areas that could be improved for clarity and conciseness, which would enhance the manuscript's overall quality.

Abstract:

Lines 11-12: The term "framework" may not be the most suitable here. The focus of the study seems more on the evaluation and comparison of models rather than the development of a framework. Perhaps rephrasing this part to "model evaluation" or a similar phrase would better capture the essence of the study.

Response #1: We thank the reviewer for the important comment and have updated the text.

Change made to the manuscript (Abstract):

This study evaluates and compares flood damage models for rice.

Comment #2: Introduction:

The introduction is quite detailed and comprehensive, but it might benefit from being more concise and directly focused on the specific research gap this study addresses.

Response #2: We very much appreciate the suggestion and have updated the introduction as follows:

1. We have moved Section 3.5.1 on the inventory of flood damage models for rice in the literature (which informed the scope of this study) to the introduction (Section 1.3). The section highlights the gaps in existing flood damage models for rice emerging from the inventory.

2. We revised the objectives in the introduction (Section 1.4).

3. We added a paragraph to the objectives in Section 1.4, which describes challenges that the study aims to address (see Response #4).

4. We have also added a graphical abstract to provide additional clarity and reader orientation.

Change made to the manuscript (Section 1.4):

The objective of this study is to conceptualize and implement a four-step methodological framework for advancing flood damage models for rice, which is also applicable to other crops. The framework supports flood damage modeling for crops that integrates machine learning approaches in model development and validates models across regions.

Comment #3: Since the study combines empirical data with machine learning techniques, it would be helpful to emphasize how this combination addresses gaps in existing research and highlight the specific problems this study aims to solve.

Response #3: We thank the reviewer for the valuable suggestion and have added a paragraph to Section "1.4 Research Contributions and Scope".

Change made to the manuscript (Section 1.4):

The methodology aims to advance flood damage models for rice crop losses. Two major challenges exist: first, a lack of uncertainty estimation, and second, transferability challenges. Empirical data (reported from farm owners) provide real-world observations of hazard, exposure, and damage for calibrating and validating flood damage models. Machine-learning models use empirical data to learn complex, multi-level relationships between flood characteristics and resulting losses, often outperforming traditional stage-damage functions. Combining empirical data with machine learning-based probabilistic models has enabled transferable and reliable flood damage predictions (Rözer et al., 2019; Sairam et al., 2020).

Comment #4: Table Clarifications:

Table 1: The distinction between "data collection period" and "flood event covered" is not entirely clear. It would be helpful to clarify the difference between these two categories for better reader understanding.

Response #4: We thank the reviewer for this comment. We have revised the two column titles in Table 1, which are now called “household survey period” and “Flood events covered by the survey questionnaire.”

Comment #5: Table 2: The definition of "flood duration" (1-100 days) could be explained in more detail. Is this range based on specific criteria or events? A brief clarification would be beneficial.

Response #5: We have revised the title of Table 2 for more clarity.

Change made to the manuscript (Section 2.2):

Table 1: Overview of predictor and response variables used for flood damage model fitting. The variables describe flood events reported in the household surveys conducted in Thailand and Myanmar described in Table 1. The ranges indicate the minimum and maximum values reported in the household surveys.

Comment #6: Model Development:

This section would benefit from a more detailed explanation of the model development process. Adding a flowchart or a clearer step-by-step description could help readers follow the methodology more easily.

Response #6: Thank you for this suggestion. We have added a Figure presenting the model development and validation steps.

Change made to the manuscript (Section 2.2):

Figure 3 provides an overview of the developed models and validation steps.

Comment #7: It seems that three models—regression, Bayesian regression, and Random Forest—are being compared. A bit more detail on each model’s methodology and how they were applied to the data would improve understanding.

Response #7: Correct, linear regression, Bayesian regression (univariable and multivariable versions), and RF were compared. We have added details on each model in the text of “Section 2.2 Model development” and in Table 3. The new Figure 3 presents which data (subsets) were used for each model. It also visualizes each step of the model validation.

Comment #8: Table 3: It might be helpful to briefly explain how the Univariable Bayesian Regression, Multivariable Bayesian Regression, and Random Forest models calculate relative yield loss, rather than placing all the details in the supplementary materials. This would improve the readability and flow of the manuscript.

Response #8: We agree with the comment and added additional details and formulas in Table 3.

Reviewer 3:

We very much appreciate the time and effort that the reviewer dedicated to providing feedback on our manuscript and are grateful for the comments and suggestions for improvements. We have incorporated the suggestions. Below, we provide point-by-point responses to the reviewer's comments.

Comment #1: This study introduced a framework named CROPDAM-X for developing and evaluating flood damage functions for crops and applied this framework to rice yield loss estimates in Thailand and Myanmar. This framework also included comprehensive review of the state-of-art flood damage models for rice and provided practical guidance for further applications. Results showed that data-driven models like Random Forest achieved the highest accuracy, while challenges remained when these models were transferred to different areas. Overall, this study is helpful for flood damage estimates in the agriculture sector. However, I still have several concerns and suggestions for improving the current work.

1) Random Forest is just one of the commonly used machine learning models. Could you justify why Random Forest rather than the other machine learning models/algorithms was employed in this study?

Response #1: Random Forest was selected because it has performed well in past flood-loss studies (e.g., Merz et al., 2013, Wagenaar et al., 2017) and is suitable for representing non-linear multivariable relationships using relatively small datasets. We have updated the paper to provide this rationale.

Change made to the manuscript (Section 2.2):

Random Forest was selected because it has performed well in past flood-loss studies (e.g., Merz et al., 2013, Wagenaar et al., 2017) and is suitable for representing non-linear multivariable relationships using relatively small datasets

Comment #2: Line 25: Please provide the data source for estimated losses due to extreme events.

Response #2: The source is the same as referenced in the following sentence. We have revised the text accordingly.

Change made to the manuscript (Section 1.1):

Extreme events have caused estimated losses of USD 3.8 trillion in the agricultural sector over the past three decades (1991-2021) (FAO 2023). This is equivalent to an average annual loss of about USD 123 billion or 5% of the global agricultural GDP (FAO 2023).

Comment #3: Lines 109-110: The proposed framework was named as "CROPDAM-X", in which "DAM" represents "DAMage". "DAM" may be interpreted as the hydraulic structure, dams, which is a little confusing.

Response #3: We very much appreciate the valuable comment and have revised the name to Crop-Loss-X.

Comment #4: Fig. 1: It is suggested to avoid the acronyms, e.g., MAE, MBE, and CRPS, (which are not explained until in the following sections) in the figure.

Response #4: We thank the reviewer for the valuable comment and have spelled out the acronyms in the revised Figure 2 (previously Figure 1).

Comment #5: Line 131: Two variables related flood characteristics were used to develop the models. Is it possible to incorporate the output from existing flood models so that more hydraulic variables can be used for the damage model for crops?

Response #5: We thank you for the suggestion. Incorporating additional hydraulic variables would be valuable in future work. At present, spatially and temporally complete flood model outputs for Thailand and Myanmar, including those hydraulic variables, are not available for the survey locations or historical event sets. We have incorporated the suggestion in the conclusion, where we describe opportunities for future research.

Change made to the manuscript (Section 4):

Expanding the datasets to include more variables and a broader spectrum of flood characteristics (from in-situ measurements or flood models) would provide insights into additional damage processes. This could improve performance as well as transferability and address open questions regarding the variable importance in the RF models.

Comment #6: Table 3: Are there any references or evidence showing that the minimum damageable flood depth is 2cm? What is h_saturation? Why was the root square of water depth instead of the other transformations used in the linear regression equation?

Response #6: We thank the reviewer for the questions. 2cm was the lowest reported flood depth for which damage of rice plants was reported in the household survey. It is the minimum value of water depth in the dataset. We have added the value of $h_{saturation}$ = 314 cm to Table 3.

The square root of water depth was used in the linear regression in line with previous studies on flood damage models for buildings, where this function is commonly used as a reference function in studies that assess the added value of more complex models. We have added a sentence under “2.2 Model development”.

Change made to the manuscript (Section 2.2):

In line with similar studies (Merz et al., 2013; Schoppa et al., 2020; Schröter et al., 2014; Wagenaar et al., 2017), the linear regression uses the square root of water depth, which is commonly used as a reference damage model for assessing the value of additional model complexity.

Comment #7: Line 188: What kernel functions are used for the density estimation?

Response #7: We thank the reviewer for the question and added a response in Section 3.1, where we describe Fig. 4.

Change made to the manuscript (Section 3.1):

The kernel density estimations were computed using the default Gaussian kernel in the R stats::density() function to produce the violin plot visualization.

Comment #8: Section 3.2.2: The performance comparison of between the ramp functions and the proposed models is based on the median or mean of the evaluation metrics? Line 235: What does it mean that the MAE for RF and BRM covered that of the ramp functions? The IQR or whiskers covered 23%?

Response #8: We thank you for the valuable questions. The thick line in the boxplots indicates the median. No CV was conducted for the ramp functions; the lines (red line and blue line) indicate the performance metrics from a validation that used all observations from the Shrestha et al. Myanmar dataset (solid red line) for testing. Figure 5 does not show the mean, but the boxplots show the median (thick line in the boxplot), IQR (min and max of the box) and 1.5x IQR (whiskers). The mean is provided in Table S7. We have added a sentence in Section 3.2.2 to clarify this.

Change made to the manuscript (Section 3.2.2):

To enable a more direct comparison with the ramp functions, we also trained all models using only the Shrestha et al. (2021) dataset. The findings from comparing the ramp functions with the mean performance metrics obtained from a 10-fold CV are summarized below and presented in detail in the Supplementary Information (Table S7). In this restricted setting:

- The MAE for RF and BRM converged with that of the ramp functions (around 23%).

Comment #9: The caption for Fig. 4 may not be correct.

Response #9: We thank the reviewer for spotting this error and corrected the title of Fig. 6 (which was originally Fig. 4).

Change made to the manuscript (Section 3.3):

Fig. 6: Box plots indicating the predictor importance for the generalized and localized RF models. Each boxplot shows the median (thick horizontal line), the interquartile range (IQR), the most extreme values within 1.5xIQR (whiskers), and outliers (black points).

Comment #10: Line 289: Could you quantify how skewed the training data is and how did that affect the model performance?

Response #10: Figure 4 shows the skewness in the training data. Loss ratios, particularly in Thailand, are a bit skewed to the higher end, which could reduce performance for lower losses. This skewness was partially caused by our focus on larger loss events in the questionnaire for Thailand. It does not contain zero-loss observations. The Myanmar data contains zero-loss observations; it is less skewed towards high loss than the Thailand data, but has a narrower range of flood duration and water depth conditions.

Comment #11: Section 3.5: It would be better to place Section 3.5.1 in the first few sections of this manuscript. Also, please note that there is no one model that fits all, so the ensemble model method might be a better option given various uncertainty sources (see the reference below).

Reference: Huang, T., & Merwade, V. (2023). Uncertainty analysis and quantification in flood insurance rate maps using Bayesian model averaging and hierarchical BMA. Journal of Hydrologic Engineering, 28(2), 04022038.

Response #11: We thank the reviewer for the suggestion and have moved Section 3.5.1 on the inventory of flood damage models for rice to the introduction (now, Section 1.3). We have added a suggestion to assess the performance achieved using the ensemble model methods in the Conclusion as an opportunity for future research, citing the paper by Huang and Merwadae (2023).

Change made to the manuscript (Section 4):

Future studies could investigate whether model ensembles, e.g. Bayesian model averaging (Huang and Merwade, 2023), yield higher performance scores than single flood damage models for rice. While we apply CROPDAM-X to rice, it could be applied to other crops and regions in the future.

Comment #12: Fig. 5 is a stacked bar chart, which makes it difficult to compare the difference in terms of the share of each category.

Response #12: We thank the reviewer for the comment and have added percentages to the stacked bar chart (which is now included in the introduction), hoping that this facilitates the readability of the share of each category.

Comment #13: Line 409: Please explain the metrics values are referred to the median or the mean.

Response #13: We thank the reviewer for the valuable comment. The reported performance metrics were mean performance metrics. In line with the suggestion by review 4 that the conclusion was too detailed, we removed the detailed performance metrics. The performance metrics are provided in the results.

Change made to the manuscript (Section 4):

Our findings highlight that, among the generalized models, the Random Forest (RF) model performs best, followed by the multivariate Bayesian Regression Model (BRM).

Reviewer 4:

We very much appreciate the time and effort that the reviewer dedicated to providing feedback on our manuscript and are grateful for the comments and suggestions for improvements. We have incorporated the suggestions. Below, we provide point-by-point responses to the reviewer's comments.

Comment #1: The authors of the article propose a method for comparing models aimed at estimating variations in rice yields based on survey data following flood events in several case studies. The scientific stakes are high, as rice is a crop that is heavily impacted in Asia. While the structure of the article is clear, further clarification of the stated objectives and real contributions of the article is needed as well as a discussion.

1. Introduction

The authors would benefit from presenting the issues in terms of model performance and the issues in terms of modelling flood-related agricultural damage separately. A literature review on the issues of modelling flood-related agricultural damage already exists (Bremond et al, 2013). Why not focus on the specific modelling of yield variations in rice and target its specific dimensions?

In my view, the scientific contribution of the article lies in the comparison of different modelling approaches and their performances. This part needs to be consolidated in the state of the art.

In the Research contributions section, the objectives stated are not really those presented in the article:

1. an inventory of flood damage models for agriculture → This is an inventory of models for rice

2. the article does propose a four-step methodological framework, but it is only applied to rice cultivation.

Response #1: We thank the reviewer for the thorough reading and have revised Section 1.3 to clarify and better align objectives with research contributions.

Change made to the manuscript (Section 1.3):

The objective of this study is to conceptualize and implement a four-step methodological framework for advancing flood damage models for rice, which is also applicable to other crops. The framework supports flood damage modeling for crops that integrates machine learning approaches in model development and validates models across regions.

The methodology aims to advance flood damage models for rice crop losses. Two major challenges exist: first, a lack of uncertainty estimation, and second, transferability challenges. Empirical data (reported from farm owners) provide real-world observations of hazard, exposure, and damage for calibrating and validating flood damage models. Machine-learning models use empirical data to learn complex, multi-level relationships between flood characteristics and resulting losses, often outperforming traditional stage-damage functions. Combining empirical data with machine learning-based probabilistic models has enabled transferable and reliable flood damage predictions (Rözer et al., 2019; Sairam et al., 2020).

Comment #2: The name of the CROPDAM-X methodological approach seems to me to echo the floodam.agri method (https://floodam.org/floodam.agri/). I encourage the authors to look at the specific features of their approach compared to this existing one.

Response #2: Thank you for pointing us to this highly relevant work. floodam.agri is a useful framework for assessing flood losses in the agricultural sector. Compared to our approach, it relies more on expert judgement, and it is more process-based than our purely data-driven approach. We had referenced the paper only in the section "1.2 Spatial transferability of flood damage models," but have now added a reference to the paper in section 1.4 "Research contributions and scope", which compares our presented framework with the floodam.agri method.

Change made to the manuscript (Section 1.4):

Compared to an existing methodological framework for developing process-based flood damage models that rely on expert judgement (Brémond et al., 2022), the methodological framework presented in this study uses a purely data-driven approach.

*Comment #3: **2. Methodology***

The experimental design does not seem clear enough to me to judge the relevance of the analyses carried out subsequently. What data were used for modelling ? On which case studies? How many simulations? It seems to me that a diagram or table should clarify these aspects.

Similarly, the presentation of the survey data on which all the performance analyses are based is not detailed enough to understand the relevance of the model transferability: What data was collected in the surveys? Yields? Variation in yield? If so, what was the reference yield for the different case studies? When did the floods occur in the case studies? …

Response #3.1: We thank the reviewer for the valuable comments. Section 2.2 "Model development" summarizes which data were used for each of the model categories. In order to provide further clarity and reader guidance, we have also added a new figure (Fig. 3) that summarizes the entire modeling framework and validation steps. The box on model development presents the data used. The data is further presented in Tables 1 and 2. The data consists of household survey data from Myanmar from the literature (Win et al. 2018 and Shrestha et al. 2021) and household survey data from Thailand collected by the authors. In addition to the reference in the data availability statement, we have revised Section "2.1 Data for the model development" to refer to the relevant section in the Supplementary Material that describes the survey methodology.

Change made to the manuscript (Section 2.1):

In addition to the secondary datasets, we conducted a household survey among farmers in Northeast Thailand to collect flood damage data concerning rice. The survey data collected in Thailand are comparable to those from Myanmar. We interviewed 584 households (20% of the 2,904 total) in the Lower Songkhram River Basin in March 2023, exceeding the minimum sample size for a 95% confidence level (see a map in Fig. S1 and Tables S1-S2 for details on the data collection). The methodology of the household survey conducted from March 11-28, 2023, is described in detail in the Supplementary Information (1.2). The data used for the model development is publicly available (Bill-Weilandt et al., 2025).

Response #3.2: Three datasets were used and made available (see reference in the Data Availability Statement):

1.) Win et al. 2018 provided the regular yield and the reduced yield due to the flood event (each in kg/ha). The values of the regular yield range from 1547 kg/ha to 4126 kg/ha. The minimum and maximum values can be calculated based on the published dataset. We have calculated the relative yield loss with the formula: (regular yield) / (yield after flood event).

2.) Shrestha et al. 2021 provided the relative yield loss in percent but no absolute yield in normal years and flood years.

3.) The dataset collected by the authors (Bill-Weilandt et al. 2025) includes the yield loss ratio in percent. The Thai Office of Agricultural Economics has published the following average productivity for wet season rice production in Nakhon Phanom Province, where the survey was conducted.

Reference: Thai Office of Agricultural Economics Zone 3 (สำนักงานเศรษฐกิจการเกษตรที่ 3): Dashboard of important agricultural products: Yield per rai in kg in Nakhon Phanom. In-season rice and off-season rice, 2025.

year	kg_per_rai	kg_per_ha
2017/18	370	2312.5
2018/19	349	2181.25
2019/20	350	2187.5
2020/21	350	2187.5
2021/22	356	2225
2022/23	351	2193.75
2023/24	350	2187.5

The response variable of the developed models is relative yield loss (in percent). We have therefore limited the information in the paper to the observed relative yield loss and provided the range of the values in the combined dataset in Table 2. Fig. 4 presents the distribution of each variable per region.

*Comment #4: **3. Results and discussion***

It would be appropriate to present the model outputs for the various case studies before comparing their performance.

Response #4: Thank you for this suggestion. We added a figure showing predictions and observations in the Supplementary Information. The figure was placed in the Supplementary Information to keep the main results section focused and readable, and a reference to it has been inserted at the start of Section 3.2.1, so readers encounter model outputs before performance metrics.

Change made to the Supplementary Information:

Suppl. Fig. 1: Predictions (for the deterministic model) and predictive densities (for probabilistic models) of relative yield loss for five randomly sampled farms. The figure shows predictions and observations generated with the generalized Random Forest (RF) model, the Bayesian Regression Model (BRM), and the probabilistic and deterministic stage-damage functions (SDF). The plots were created with the generalized models trained on all the data (Leave-Nothing-Out). Dashed orange lines indicate the observed relative yield loss. Solid blue lines indicate the prediction of the deterministic model and the mean of the predicted distributions of the probabilistic models.

Comment #5: The results presented must be consistent with the stated objectives. It is unclear whether the objective is to compare performance or to inventory and discuss the performances of the various approaches. This needs to be clarified.

Response #5: We thank the reviewer for this important comment and have revised Section 1.4 on research contributions and scope of the study to clarify the objectives of the study. We also revised the conclusion, providing the broader contributions of the study in the first paragraph.

Comment #6: A section discussing the results is missing.

Response #6: In order to avoid repeating results in the discussion section, we have included combined results and discussion in Section 3. Here, we present the discussion directly after presenting the results for each component of the study.

Comment #7: 4. Conclusions

The conclusion needs to be improved. However, this requires first clarifying the main focus of the article.

Response #7: Thank you for the comment. We have improved the conclusions in the context of improved objectives and overall focus of the study (please see also the changes in Section 1.4 concerning the scope of the study).