Response to RC1

Overall Assessment

- The manuscript presents a framework for assessing potential future typhoon-induced disasters by coupling high-resolution meteorological simulations with river discharge and storm-surge models, applied to Super Typhoon Hagibis (2019) as a case study. The framework combines three typhoon models, three river models, and two storm-surge models with multi–initial-condition ensembles to represent structural model uncertainty and internal variability. The results suggest intensification of the typhoon under warming, with increased precipitation, stronger near-surface winds, and lower central pressure, accompanied by higher river discharges in many eastern-Japan basins and larger storm surges. While the framework is conceptually valuable and the experiment design is ambitious, in its current form the manuscript does not yet meet the requirements for publication. The concerns outlined below, particularly regarding robustness, interpretation, and implementation of the framework, should be addressed.

Response:

We thank the editor and the reviewers for their careful reading and constructive comments. We have revised the manuscript accordingly. All changes are indicated in the revised manuscript.

General Comments

- 1. The paper primarily focuses on Super Typhoon Hagibis (2019) as a case study, which may limit generalizing the findings to other typhoons or extreme weather events. Further research is needed to validate the framework across multiple events and regions. How well do the results generalize to other typhoons or extreme weather events beyond Super Typhoon Hagibis (2019)? Have you tested the framework on other events, and if so, what were the findings?

Response:

Using the same ensemble experiment framework, we also conducted additional experiments for Typhoon Jebi (2018). Similar to the case of Typhoon Hagibis, we found an intensification of both the typhoon and the associated storm surge in the warming-conditioned experiments. However, because these results have not yet been published in a peer-reviewed paper, we do not use them as formal evidence in the manuscript.

We also note that many previous studies have reported the intensification of typhoons under global warming; therefore, the results of the present study are consistent with the existing literature. Importantly, the objective of this study is not to quantify the general influence of

global warming on typhoons in a probabilistic sense. Rather, we adopt an event-based storyline approach to examine climate change impacts on an individual tropical cyclone and to develop a framework for conducting such event-specific assessments. This approach enables detailed analysis of storm evolution and hazard-relevant processes, while it does not aim to represent probabilistic changes in future occurrence or track statistics.
To obtain a comprehensive picture of tropical cyclone risk under future climate conditions, the Ev-SL framework should be complemented by probabilistic approaches (e.g., large-ensemble simulations) that can address changes in frequency and tracks. We have clarified this positioning and scope of the present study in line 379-384.

- 2. The framework proposed in this study assumes a linear coupling of the GSM data and the difference component of the climate change of air temperature and sea surface temperature between the +2 or +4 K future and 2023. This assumption may oversimplify the actual climate change impacts and should be further validated. How does this assumption impact the accuracy of the results, and are there alternative methods or models that could be used for a more sophisticated coupling?

Response:
The reviewer raises an important point regarding the assumption of linear coupling in the PGW framework. This approach indeed simplifies the representation of climate change signals and may not fully capture nonlinear interactions in the climate system. However, as described in this study, our objective is not to reproduce the full probabilistic range of future climate variability, but rather to conduct a stress-testing analysis of a specific high-impact event under different warming levels. In this context, the PGW method provides a practical and widely used approach for isolating thermodynamic effects while maintaining the dynamical structure of the event. In lines 327–337, we summarize the limitations and advantages of the PGW approach.

3. The storm surge and hydrological models are used separately rather than within a fully coupled framework, which restricts the representation of dynamical interactions within the typhoon-rainfall-runoff-surge system. In reality, storm structure, precipitation, river discharge, and coastal water levels evolve in a tightly linked way, and this coupling is not explicitly represented when the two impact models are run independently. It also remains unclear how the water levels (or flood indicators) produced by the two models are combined; flooding hazards should be assessed using both sources of information with particular care, especially because coastal surge levels are influenced not only by meteorological forcing but also by river inflow at river mouths.

Response:
As the reviewer points out, river water levels change due to storm surges in estuary areas. However, since this study evaluates changes in river flow upstream of the estuary, this impact is not considered. River water levels also affect storm surge levels near the estuary, but this influence is negligible and therefore disregarded. The text in lines 177-179 has been revised according to the reviewer's comments.

4. When comparing Fig. 4(b) with Fig. S2(b), the maximum wind speed does not increase under warming for the NHRCM and WRF simulations, in contrast to CReSS and to the warming-induced intensification indicated by changes in central pressure and precipitation. Consequently, the authors' statement that "all three of these results indicate the strengthening of typhoons due to global warming, suggesting that storm surges and flooding will also become more hazardous accordingly" (lines 200–201) is not fully justified as written. This inconsistency is counter-intuitive and potentially important, yet it is not explored or discussed in sufficient detail in the current version of the paper.

Response:
The 15 simulations shown in Figs. 4 and S2 consist of five initial-condition ensemble members simulated with three different convection-permitting models (CPMs). In the figures, cross marks indicate individual simulations, while solid circles denote the ensemble means calculated across the five initial-condition members for each CPM. For each CPM, the ensemble-mean maximum wind speed increases under warming. The quantitative values reported in the manuscript (e.g., 2.97 m s^{-1} in +4K and 2.75 m s^{-1} in +2K) correspond to the multi-model mean, obtained by averaging the ensemble means of the three CPMs. We acknowledge that this hierarchical averaging was not sufficiently clear in the original manuscript. To avoid misunderstanding, we have revised the figure captions and relevant text to explicitly describe how the ensemble means and multi-model means are defined in lines 246-247.

5. The paper does not extensively discuss the computational requirements and potential limitations of the framework, such as the resource-intensive nature of running multiple models and ensemble experiments. What are the computational requirements for running the framework, and how do they scale with the number of models and ensemble members? Are there strategies to optimize the computational efficiency of the framework?

Response:
This study was conducted as part of an approximately one-year project, during which we performed the Hagibis experiments and obtained the results described in the manuscript.

Because the computational environment was changed during the course of the project, we regret that it is difficult to provide a reliable estimate of the actual computational time and the computational resources required for the calculations.
Regarding how the computational cost scales with the number of models and ensemble members: for the models, the cost depends on the computational burden of each model introduced; for the number of ensemble members, since the experiments are conducted in parallel, the total computational cost scales linearly with the number of members

Specific Comments

1. Line 34-36: The statement about changes in extremes appears to be phrased as globally general. Please verify that it is valid for all regions or reword to reflect regional differences reported in the literature.

Response:

We agree that the original wording was overly general. In the revised manuscript in lines 33–34, we have rephrased the statement to clarify that increases in extreme events are projected in many regions, as reported in IPCC AR6 (2021, 2022).

2. Fig. 1: The conceptual framework in Fig. 1 is generally helpful, but it would benefit from clearer detail on which variables are downscaled and passed between the different models. In addition, the acronym "SV" should be explicitly defined (presumably "singular vector"), as it is used before being introduced.

Response:

In the revised manuscript, Fig. 1 has been substantially improved to explicitly indicate the atmospheric variables passed to each model component. Specifically, precipitation (Pr), surface air temperature (Ts), specific humidity (q), wind speed (Us), downward shortwave and longwave radiation (SW↓, LW↓), and cloud cover (CC) are shown as inputs to the hydrological models (RRI, 1K-DHM, and MATSIRO). For the storm surge model, the 10-m wind vector (Us, Vs) and sea level pressure (SLP) are explicitly indicated. In addition, the term "singular vector" is now written in full in Fig. 1 to avoid ambiguity. We believe these revisions significantly improve the transparency and clarity of the modeling framework.

3. Please adopt a consistent term for tables throughout the manuscript; "Table" and "Tab." are currently used interchangeably and should be unified.

Response:

We have revised the manuscript to ensure consistent terminology throughout and have

unified all references to tables as “Table.”

4. Line 107-108: Provide more detail on the criteria and procedure used to reduce the 27 ensemble members to 5. For example, thresholds on track, landfall location, timing, or intensity.

Response:

We acknowledge that the selection procedure was not sufficiently described in the original manuscript. In the revised manuscript in lines 121–127, we have clarified that the five cases were selected based on the reproducibility of a typhoon track passing through Tokyo Bay, consistent with the observed Typhoon Hagibis. Because the objective of this study is to assess compound flood risk in Tokyo Bay, maintaining track consistency relative to the bay is essential. This criterion is particularly important for storm surge simulations, as surge height is highly sensitive to the typhoon track. Even relatively small deviations in track position can substantially alter wind forcing and pressure effects, resulting in large differences in simulated surge levels.

5. Line 161-164: The authors explicitly state that they do not compare simulated river discharge with observations because the model reproducibility is “not enough,” yet they still use these simulations to infer robust relative changes in flood severity; this raises a fundamental credibility issue for the discharge results.

Response:

We acknowledge that the original wording may have created confusion regarding the credibility of the discharge simulations. In the revised manuscript in lines 207–214, we have clarified that although the simulated discharge does not perfectly reproduce observed magnitudes, the overall hydrological response is reasonably represented.

The analysis focuses on relative changes between present and warmed climates using the same model configuration and initial conditions. Under the PGW framework, systematic model biases are largely reduced when evaluating relative differences. We have revised the text to clearly distinguish between limitations in absolute reproducibility and the robustness of relative change assessment.

6. The figures are not introduced in numerical order; for example, Figure 6 is cited before Figures 2-5, and similar inconsistencies occur throughout the manuscript. This issue also applies to figures in SM.

Response:

We have carefully revised the manuscript to ensure that all figures are introduced in

numerical order according to their first appearance in the text. Specifically, we removed forward citations (e.g., to Fig. 6 in the Methods section) and revised the text to avoid referring to figures before they are introduced.
In addition, we have confirmed that each figure is explicitly referenced in the main text, including Fig. 2, which is now properly introduced. The numbering and citation order of all Supplementary Figures have also been corrected to ensure consistency.

7. In figure S4, the two panels use different color scales and ranges for the maximum tidal level deviation, which makes direct visual comparison between SuWAT and GeoClaw difficult and potentially misleading. For a fair model-to-model comparison, it would be preferable to adopt a consistent colormap and identical value range in both panels, so that the same color represents the same tidal deviation in each plot.

Response:
We agree that a consistent color scale is essential for fair model-to-model comparison. In the revised Supplementary Figure S4, we have adopted an identical colormap and value range for both SuWAT and GeoClaw panels, ensuring that the same color represents the same tidal level deviation in each plot. This modification improves visual comparability between the two models.

8. Line 243-244: The discrepancy between simulated and observed storm surge levels appears too large for a study that aims to provide robust projections, which should be preceded by strong validation. Even if the ensemble maximum matches the observation, it remains important to bring the ensemble mean closer to the observed values, and the causes of this bias should be examined and discussed more thoroughly.

Response:
In the revised manuscript, we have modified Figure 7 to explicitly highlight the ensemble member that produced the highest storm surge level (solid line). Among the present-climate simulations, this case shows the best agreement with the observations and yields the highest correlation coefficient with the observed time series. The peak surge level is also reproduced reasonably well.
Storm surge height is highly sensitive to typhoon track. In several ensemble members, relatively small deviations in track position from the observed path resulted in substantially lower surge levels in Tokyo Bay. Therefore, the ensemble mean does not necessarily coincide with the observed value, as it represents an average over realizations with varying track positions.
Figure 8 focuses on the projected change in maximum surge levels across ensemble

members. Because the objective is to assess potential changes in extreme surge under warming conditions, the analysis emphasizes the change in peak surge among physically plausible cases rather than the mean surge level. We have revised the manuscript to clarify these points.

9. Line 355: The label "(a)" is missing from the figure caption.

Response:

The label "(a)" has been added to the figure caption, and the text has been revised for clarity and consistency.

10. Line 265: The term "multi-multi ensemble" is catchy but might be confusing. It is recommended to consider using "multi model, multi–initial condition ensemble (MM–MI)" consistently, and introduce any shorthand once.

Response:

The terminology has been revised, and the abbreviation "MM–MI" is now used consistently throughout the manuscript.

11. Section 2.1: In the title, the "parent mode" should be "Parent model"?

Response:

The typo has been corrected. "Parent mode" has been revised to "parent model" in Section 2.1.

12. Many of the projected results for the +2 K scenario are placed only in the Supplementary Material. These should be moved alongside the +4 K results in the main text so that readers can more easily compare the two warming levels.

Response:

The central findings of this study are presented in Figures 6 and 8, where the projected changes under both the +2 K and +4 K warming scenarios are shown side by side. Therefore, direct comparison between the two warming levels is already possible in the main text.

For the remaining results, we intentionally retained the detailed +2 K outputs in the Supplementary Material. Including all +2 K figures in the main text would substantially increase the number of panels and obscure one of the key messages of this study, namely the spread associated with the MM–MI ensemble framework.

We believe that the current structure maintains clarity by highlighting ensemble variability in the main text while still providing complete information for both warming

levels in the Supplementary Material.

13. It seems better to provide at least one numerical example in the main text where the multi-model maximum vs single model range makes a qualitative difference for design (e.g., “Naruse River +4 K” or “Tokyo Bay surge +4 K”), with concrete magnitudes.

Response:

In the revised manuscript, we have added a specific numerical example to illustrate the practical significance of the MM–MI framework. For the Naruse River under the +4 K scenario, the MM–MI maximum indicates a 3.15-fold increase in peak discharge, whereas the expected increase derived from a single model is 1.83-fold.

This example clarifies how the multi-model maximum can lead to qualitatively different implications for flood management compared to a single-model projection.