A multiscale modelling framework of coastal flooding events for global to local flood hazard assessments

Responses to reviewers:

Reviewer #2:

This manuscript presents a methodology for evaluating flooding at high resolution by coupling three models: GTSM, Delft3D, and SFINCs. These models are among the latest and most robust developments in hydrodynamic modeling. On one hand, it is necessary to develop robust methodologies to assess coastal flooding, taking into account different types of forcings such as tropical cyclones (TCs) and extratropical cyclones (ETCs). On the other hand, this study tries to emphasize the importance of increasing both temporal and spatial resolution as well as enhancing hydrodynamic flood modeling.

Therefore, this research and development could be a valuable contribution to the scientific community. However, I believe this study fails to convincingly demonstrate that refining, downscaling, and dynamic flood modeling significantly improve flood hazard assessment. Several aspects need clarification, and a clear message about your results, conclusions, or recommendations for performing a flood risk assessment has not been adequately addressed.

We are pleased that the reviewer recognises the potential scientific contribution of our manuscript. In response to the detailed and constructive feedback, we have made substantial revisions to the manuscript.

We have rewritten the introduction to clearly outline the limitations addressed by our study, presenting a modelling framework that is easy to use and flexible. To do so, we have conducted a sensitivity analysis to examine the impacts of model resolution on the simulation of total water levels and flood extents. We have also included validation of the total water levels and flood extents into the manuscript to strengthen the foundation of our modelling framework. It is important to note that from the model validation we cannot conclude that refinements always lead to an improved model performance. Instead, each refinement is case specific and therefore the user of the modelling framework should decide depending on the type of storm and location of the analysis which refinement is most adequate.

We believe that these amendments have greatly enhanced the manuscript and now we better answer the question how downscaling affects the model performance. We thank the reviewer for the time and effort dedicated to reviewing our work.

The weakest part of your study is the sensitivity analysis of the model configurations.

First, concerning the organization, it would be advisable to assign a nomenclature to each configuration to aid in comparisons and analysis. For example, the default configuration and the refined temporal and spatial output could be assigned the same letter with different numbering (since they all originate from the same model, with the same forcing and bathymetry). The fully refined configuration should have another letter, because although it combines higher temporal and spatial resolution, the fact that the GTSM simulations use more current and detailed bathymetry (despite Europe EDMOnet being used

with the same resolution as GEBCO2014) introduces another distinct element that can significantly affect the results. Finally, the dynamic downscaling (a nomenclature related to the global configuration, to which the fully refined configuration is nested), should not be directly compared with the default configuration without analyzing the effect of the previous factors.

This distinction between configurations should also be used for comparing results at the storm surge modeling and hydrodynamic modeling levels. This way, the analyses would be more orderly, identifying how each factor influences the outcomes, leading to more conclusive comparisons and differences.

In section 3.1 (multiscale storm surge modeling), I would divide it into three subsections: the first analyzing the effect of higher resolution on the maximum water level value (results shown in Figure 5 b, e, and h), the second identifying the effect of bathymetry changes (moving from GEBCO2014 to GEBCO2023, results shown in Figure 4), once the time resolution effect is identified, and the third extracting the added value of dynamic downscaling (once the effects of time resolution and bathymetry changes are understood).

Thank you for your valuable suggestions. In response to these suggestions, we have provided a nomenclature for each model configurations, distinguishing between model configurations that are or modify the global GTSM model (G), and model configurations that include a nesting of a locally refined model (N). Furthermore, for a more stepwise sensitivity analysis of the dynamic downscaling, we have separated the process into two model configurations: first, grid refinement alone retaining the old GEBCO2019 bathymetry (N1), followed by a dynamic downscaling with updated bathymetry interpolated to GEBCO2023 (N2). We chose this approach, rather than starting with the updated bathymetry alone and then applying grid refinement, because the latter could be automated in Python, while the former required manual intervention. This stepwise approach allows to isolate the effects of the grid refinement from those of updating the bathymetry. Furthermore, as suggested by the reviewer we have split the section into subsections that identify the effects of higher resolution, changed in grid refinement and bathymetry.

We have updated the manuscript as follows:

> *(Lines 206 – 220): "Using the MOSAIC modelling framework, we analyse the effects of refining the resolution of GTSM on the simulated water levels and assess how these propagate into the results for the flood hazard simulated by SFINCS. As described in Table 1, we categorise model configurations in two distinct groups. The first group, which contains the global model configurations (G), includes the default model configuration (G1) and configurations that modify only the global GTSM model (G2 and G3). In this group, the refinements applied are: (1) the temporal output resolution, which is different than the implicitly calculated simulation timestep of GTSM, is refined from 1-hourly to 10-minute, allowing to capture more changes in water levels, including the peaks of the water levels (G2); and (2) the spatial output resolution is refined from locations along the coast every ~5 km to ~2 km, providing more coastal boundary conditions for the hydrodynamic flood hazard model (G3). The second group, which contains the nested model configurations (N), includes those model configurations that use a nested local model within the global model GTSM by performing dynamic downscaling. These model configurations include: (1) the nesting of local high-resolution models with refined grids into GTSM (N1); and (2) the nesting of local high-resolution models with refined grids and updated*

bathymetry into GTSM (N2). Finally, we evaluate the combined effects of all these refinements through the "fully refined" configuration (N3), which integrates both the enhanced temporal and spatial resolutions as well as the nested high-resolution models and updated bathymetry. The validation of GTSM and SFINCS shows sufficient performance for all the model configurations from Table 1 and Fig. 7 (see Table A1 and Figs. A2 and A3)."

We have also added the nomenclature in the table below:

Table 1. GTSM model configurations used in the sensitivity analysis.

Model configuration	Nomenclature	GTSM grid resolution	Bathymetry	Spatial output resolution	Temporal output resolution
Default configuration	G1	~25 to 2.5/1.25km	GEBCO2019*	Original (~5 km)	1h
Refined temporal output resolution	G2	~25 to 2.5/1.25km	GEBCO2019*	Original (~5 km)	10min
Refined spatial output	G3	~25 to 2.5/1.25km	GEBCO2019*	Refined (~2 km)	1h
Dynamic downscaling (Refined grid)	N1	~25 to 0.45km	GEBCO2019*	Original (~5 km)	1h**
Dynamic downscaling (Refined grid + Updated bathymetry)	N2	~25 to 0.45km	GEBCO2023	Original (~5 km)	1h**
Fully refined configuration	N3	~25 to 0.45km	GEBCO2023	Refined (~2 km)	10min**

** EMODnet2018 for Europe*

***For the model configurations N1, N2 and N3, the temporal output resolution is also the temporal resolution of the coupling between GTSM and the local high-resolution model.*

We have added a dedicated section in the results to analyse these effects. Subsequently, we have examined the impact of updating to a new bathymetry within the refined grid. In the results we have also interpreted why changes in the model configuration result in higher or lower water levels. These updates are now reflected in the manuscript as follows:

(Lines 258 – 294):

"3.1.2 Effects of dynamic downscaling with original bathymetry on total water levels

Figure 8 panels c, g, k show that the model configuration N1 results in significant changes in water levels for all case studies. The largest differences occur along the coasts, where the largest changes in model grid size resolution occur. For TC Irma (Fig. 8 panel c), the nesting of a local model at high-resolution with GEBCO2019 results in maximum water levels that are up to 0.3 m higher than G1 in the southwest of Florida, and up to 0.1 m lower in the southwest. These changes are caused by the refined grid resolution in those regions in comparison to G1, which allows us to better resolve complex topography around the barrier islands. Water levels for nine tide gauge stations along the coast indicate that while G1 underestimates the peak of TC Irma

in most locations (Fig. A2, all stations but station 7), N1 simulates on average higher peaks, resulting sometimes in overestimations (Fig. A2, station 9). Additionally, the performance of N1 is slightly better than G1 for six tide gauge stations (stations 1-6), as reflected in Table A1, which shows lower RMSE values. However, for stations 7-9, G1 shows slightly higher RMSE and Pearson's correlation. For TC Haiyan (Fig. 8 panel g), the differences in maximum water levels are up to 1 m higher than G1 near the landfall regions. These differences occur due to the refinement of the grid from 2.5 km to 45 m, which results in a significant increase in the number of model grid cells that define regions of shallow bathymetry, especially around the bay near Tacloban, resulting in a more detailed representation of water levels in that region. Thanks to the increase on grid cells, the strait north of Tacloban for N1 is defined with multiple grid cells in comparison to the two grid cell width of G1 (see Fig. A6). Therefore, in that region N1 allows us to better resolve the topography of the region, and water can travel more easily northwards. For ETC Xynthia (Fig. 8 panel k), the water levels from the nested local model at high-resolution are overall lower than water levels for the G1. Near La Rochelle, those water levels are up to 0.2 m lower. When comparing the performance of N1 with G1 (Table A1 and Fig. A3), both model configurations can predict the timeseries pattern well, with high Pearson's correlation coefficients. Overall, the RMSE for Xynthia is similar for most tide gauge stations, except for two stations located in the mouth of estuaries (stations 3 and 6).

3.1.3 Effects of dynamic downscaling with updated bathymetry on total water levels

Figure 8 panels d, h, l show that the model configuration N2 results in relatively large changes in the water levels for all the case studies. The largest differences occur along the coasts and provide figures similar to those from N1. For TC Irma (Fig. 8 panel c), the nesting of a local model at high-resolution with updated GEBCO2023 bathymetry results in maximum water levels that are 0.3 m higher than G1 in the south of Florida. Compared to N1, model configuration N2 provides slightly higher water levels south of Florida. Those differences come from differences between GEBCO2023 and GEBCO2019 in the region. N2 shows a similar performance to G1 and N1 across nine tide gauge stations (Table A1 and Fig. A2). For TC Haiyan (Fig. 8 panels h), the differences in maximum water levels are up to 1 m higher than G1 at the landfall regions. Compared to N1, N2 provides on average higher maximum water levels, except in the bay of Tacloban where N1 presents on average higher maximum water levels. These differences come from the differences in GEBCO2019 and GEBCO2023. For ETC Xynthia (Fig. 8 panels l), the water levels from the nested local model at high-resolution with GEBCO2023 are lower overall than water levels for G1. Compared to N1, the model configuration N2 provides a similar pattern of water level decrease, however, the maximum water level reduction compared to G1 is slightly less than for N1. The performance of N2, as shown in Table A1 and Fig. A3, is comparable to that of G1 and N2, except at two tide gauge stations (station 3 and 6) where GEBCO2023 does not accurately capture the bathymetry of the river channels in the estuaries. In contrast, EMODNET2018, the bathymetry used in model configurations N1 and N3, better resolves these details (see Fig. A7).

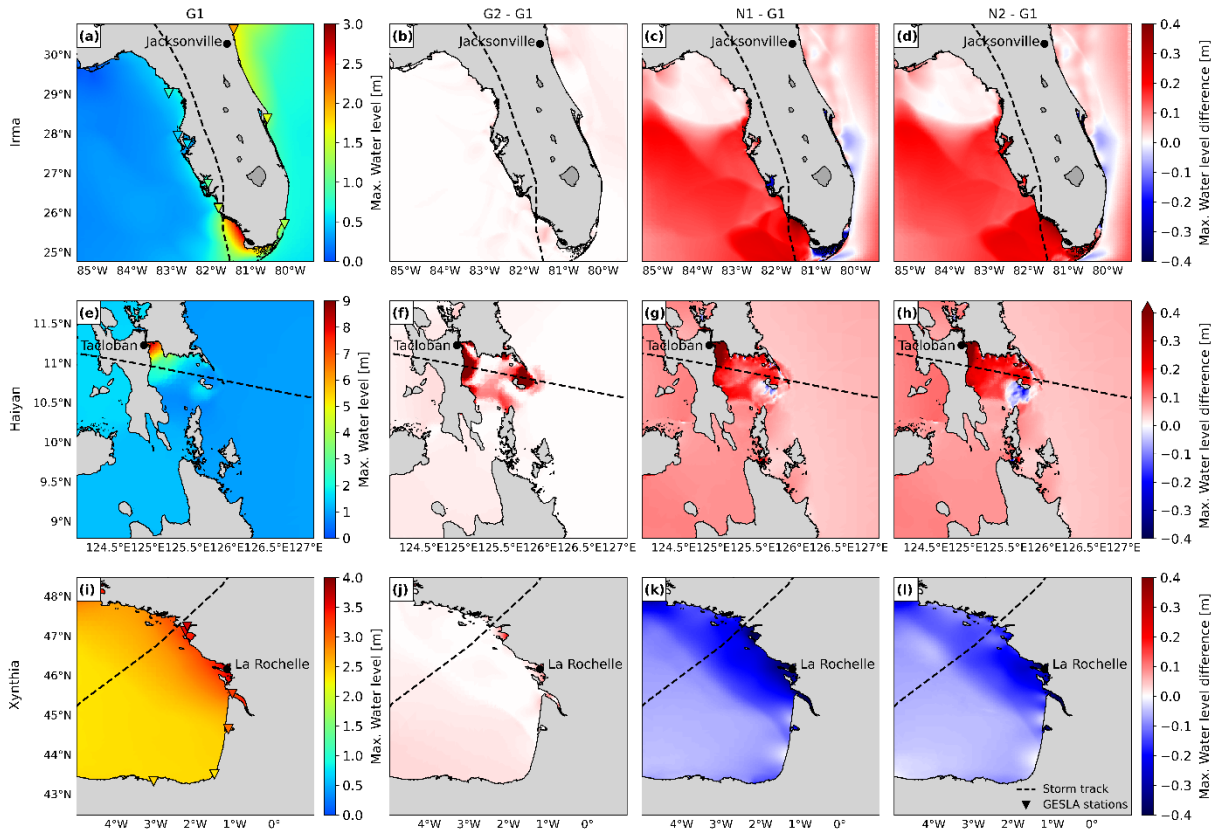

Figure 8. Maximum water levels for the three case studies for G1 (panels a, e, i). Difference between the maximum water level for each specific model configuration (see Table 1) and G1. Panels a, e, i show observed maximum water level from tide gauge stations of GESLA. Difference in water levels for G2 (panels b, f, j), N1 (panels c, g, k) and N2 (panels d, h, l)."

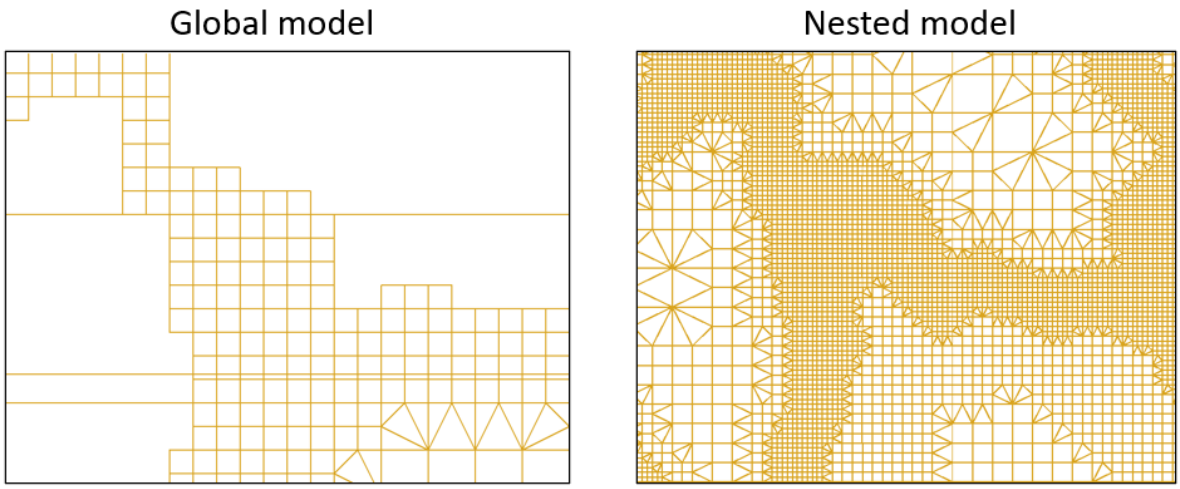

Figure A6. Close look at the unstructured grids of the global GTSM model with a grid resolution up to 2.5 km along the coast (left) and the nested grid of dynamic downscaling with a grid resolution up to 0.45 km along the coast (right), for case study Haiyan.

(Lines 332 -351):

*"**3.2.2 Effects of dynamic downscaling with original bathymetry on flood depths***

Figure 10 panels d, i , n show that the model configuration N1 results in significant changes in the flood depths for all the case studies. For TC Irma (Fig. 10 panel d), model configuration N1 leads to slightly higher water levels in comparison to G1. Consequently, the resulting flood depths are also larger and are more than 0.2 m above those of G1. Maximum water levels for TC Haiyan (Fig. 10 panel i) are generally higher along the bay of Tacloban when applying dynamic downscaling with the original bathymetry. This results on average in higher flood depths of more than 1 m compared to G1. Finally, ETC Xynthia (Fig. 10 panel n) presents lower water levels for N1 compared to G1. Those lower water levels lead to lower flood depths across the whole model domain. For ETC Xynthia, N1 shows a lower hit rate and false-alarm ratio compared to G1, and the same critical success index (see Fig. A9).

3.2.3 Effects of dynamic downscaling with updated bathymetry on flood depths

*Figure **Error! Reference source not found.** panels e, j, o show that the model configuration N2 results in s ignificant changes in flood depths for all case studies. For TC Irma (Fig. 10 panel e), model configuration N2 compared to G1 leads to higher and lower water levels, depending on the region. Consequently, the resulting flood depths for N2 vary between 0.05 m lower to more than 0.2 m higher than G1. Maximum water levels for TC Haiyan (Fig. 10 panel j) are generally higher in the bay of Tacloban for model configuration N2 (when applying dynamic downscaling with the updated bathymetry) compared to G1. This results in larger flood depths which, in some regions, result in more than 1 m higher compared to G1. However, in the Tacloban Bay N1 results on average in higher maximum water levels than N2, which leads to lower flood depths for N2 in comparison to N1. Finally, for ETC Xynthia (Fig 10 panel o) water levels are lower for N2 compared to G1. Those lower water levels lead to lower flood depths across the whole model domain. For ETC Xynthia, N2 shows a lower hit rate and false-alarm ratio compared to G1, and the same critical success index (see Fig. A9).*

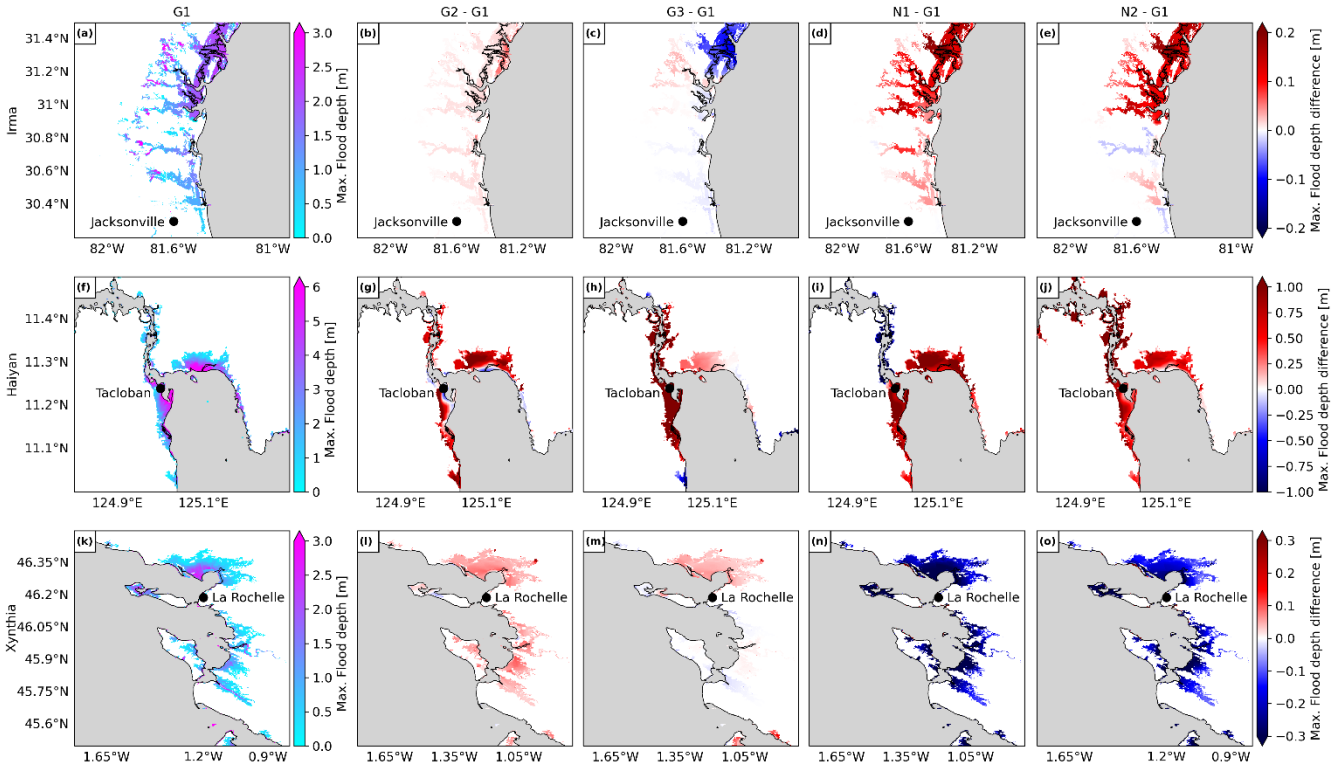

Figure 10. Panels a, f, k show the maximum flood depth for the default configuration G1, for each case study. Panels b, g, l show the difference between the maximum flood depth for the refined temporal output resolution configuration G2 and G1. Panels c, h, m show the difference between the maximum flood depth for the refined spatial output configuration G3 and G1. Panels d, i, n show the difference between the maximum flood depth for the dynamic downscaling (refined grid) configuration N1 and G1. Panels e, j, o show the difference between the maximum flood depth for the dynamic downscaling (refined grid and updated bathymetry) configuration N2 and G1."

In section 3.2 (Hydrodynamic flood modeling), I would also make a stepwise comparison, isolating the effect of different factors (spatial resolution, temporal resolution, bathymetry, and dynamical downscaling). In the first subsection, compare the default configuration with the refined temporal resolution and refined spatial scale. In the second subsection, compare the results of the fully refined configuration with the dynamical downscaling to isolate and identify this effect. The goal is to see what each factor contributes and how high-resolution modeling with SFINCs improves, once the effect of water level (as the boundary condition) is accounted for. Comparing the default configuration with the dynamical downscaling does not reveal which factor is more influential, for example, if a more detailed bathymetry in the global model is already the most determining factor.

We have done this separation in the revised manuscript and included a more stepwise analysis of the results for the dynamic downscaling. We have explained this in the previous comments, together with the changes of section 3.1.

It would be beneficial to conclude with some recommendations, as this methodology can be applied by various researchers and consultants in their flooding studies. It should be noted that the available

water level data are related to the default configuration. Therefore, it is crucial to identify the added value of each factor rather than jumping directly to the dynamical downscaling nested in the refined spatial output. Other researchers applying this methodology would start from the water level of the default configuration, even without modeling the TC forced by the Holland wind model (as in the Global Sea Level time series available at Copernicus Climate Data Store: https://doi.org/10.24381/cds.a6d42d60). How well or how it may affect using the Holland model?

Thank you for the suggestion. Adding more information and suggestions on how the modelling framework could be used by researchers does help to better understand the applications that MOSAIC could be used for. We have modified the manuscript as follows:

> *(Lines 466 – 483): "Users of MOSAIC can easily simulate storm events in any region with this the modelling framework. First, they can select the appropriate meteorological forcing. Within MOSAIC, users can choose gridded meteorological data from reanalysis datasets or climate models to simulate ETCs or TCs, provided that the data accurately captures the TC wind and pressure fields (as seen with ETC Xynthia and TC Irma in this study). Alternatively, they can select a hybrid approach that combines the Holland model with ERA5 in the background when modelling smaller TCs with rapid intensification (such as TC Haiyan in this study). Depending on the specific storm simulated and study area, users can select different model refinements. For instance, the G2 model configuration with refined temporal output resolution is suitable for rapidly intensifying storms, while nested models can resolve help resolving the topography and bathymetry in regions with complex coastlines. If the users have coastal boundary conditions available, MOSAIC can automatically generate stand-alone local high-resolution Delft3D FM models (N1 , N2, and N3 model configurations) without having to couple them with GTSM. Although uncalibrated, these model configurations demonstrate similar performance than the well-established global model GTSM (G1; see Section 3), but at a significantly lower computational cost. The hydrodynamic flood modelling part of MOSAIC offers user-defined settings as well, enabling users to, for instance, choose the most suitable DEM for their study area or implement flood protection measures through MOSAIC's HydroMT component.*
>
> *By leveraging the flexibility of MOSAIC to modify input datasets, the modelling framework can be used to study events under historical- and climate change conditions. Furthermore, taking advantage of MOSAIC's multiscale modelling approach, TC/ETC high-resolution hazard assessments can be obtained globally. When linked to impact models, MOSAIC can also be used for risk assessments."*

Another significant drawback of your study is the lack of validation with observations of water levels and flood extents. Is there any possibility to validate the steps of the MOSAIC methodology with observations of water levels or spatial flooding maps? How can we be sure that increasing spatial/temporal resolution, improving bathymetry, and employing dynamic downscaling enhance the results?

Thank you for the valuable suggestions and valid questions. We understand the concern about including relevant observations of water levels and flood extents. For this reason, we have validated the modelling framework results for the case studies for which observations are available.

We have validated the total water levels using observations from the GESLA tide gauge stations for the case studies Irma and Xynthia. Thanks to this more thorough validation, we decided to update the GTSM model version from 3 to 4.1, which has a better tidal performance. Furthermore, we decided to use ERA5 only for Irma, as this showed better results than the Holland model, which overestimated the peak of the event. For TC Haiyan, on the other hand, ERA5 alone did not capture the TC and therefore we used the Holland model combined with ERA5 in the background. We updated the manuscript as follows:

> *(Lines 101 – 104): "The meteorological forcing datasets used in this study vary per storm. For ETC Xynthia and TC Irma, we use mean sea level pressure and 10 m meridional and zonal wind components from the ERA5 re-analysis dataset at a horizontal resolution of 0.25 degrees and 1 hour temporal resolution (Hersbach et al., 2019). Because TC Haiyan is not well resolved in ERA5 (see Fig. A1), we use pressure and wind from tropical cyclone track data merged with ERA5."*

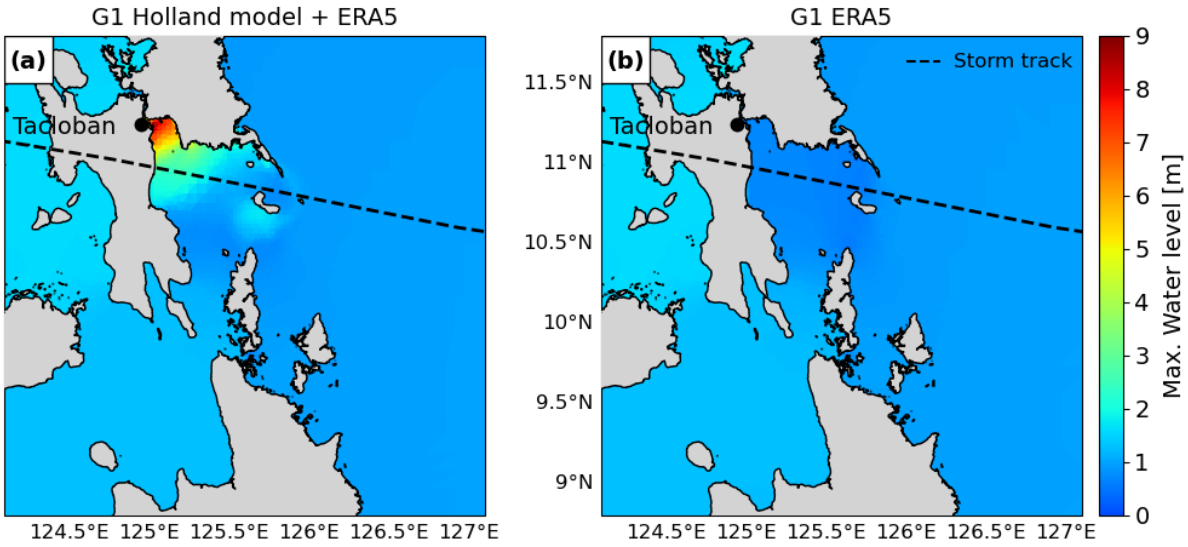

Figure A1. Maximum water levels output of GTSM, for case study Haiyan, with different meteorological forcings. Left: maximum total water levels with ERA5 as forcing. Right: Maximum total water levels with the Holland model combined with ERA5 as a forcing.

We have updated the manuscript as follows to include the validation of the total water levels:

> *(Lines 115 – 139): "MOSAIC uses GTSMv4.1 to simulate total water levels resulting from tides and storm surges, ignoring baroclinic and wave contributions. GTSM is a global depth-averaged hydrodynamic model based on Delft3d Flexible Mesh (Kernkamp et al., 2011). It has a spatially-varying resolution of 25 km deep in the ocean and 2.5 km along the coasts (1.25 km for Europe) (Dullaart et al., 2020; Muis et al., 2020) The spatially-varying resolution makes it computationally efficient for simulating water levels at large scales. The bathymetry in the model is the 15 arcseconds resolution EMODnet bathymetry dataset for Europe (Consortium EMODnet Bathymetry, 2018), and the 30 arcseconds General Bathymetric Chart of Oceans 2019 dataset for the rest of the globe (GEBCO, 2014). Tides are generated internally with tide generating forces, while storm surges originate from external forcing with pressure and fields (Section 2.1.1; Muis et al., 2020). GTSM has been successfully validated using different meteorological datasets and has been shown to provide accurate extreme sea levels (Dullaart et al., 2020; Muis et al., 2020, 2016). Version 4.1 is a calibrated version of the model with also improved parametrizations for internal tides and bottom friction coefficient (Deltares, 2021;*

Wang et al., 2022). GTSM provides as output water level timeseries over a grid in the ocean and for locations along every ~5 km of the coast.

To validate the coastal component of our modelling framework, we compare total water levels from GTSM against observed total water levels from tide gauge stations of the Global Extreme Sea Level Analysis (GESLA) dataset (Haigh et al., 2023). This comparison is made for case studies where the GTSM output locations are found nearby tide gauge stations from GESLA (see Fig. 3). GTSM output is referenced to mean sea level (MSL). We reference the GESLA water levels to the MSL by removing the annual average water level for each year, and subsequently removing the mean over the 1985-2005 period from the de-trended time series. To assess the accuracy of GTSM, we calculate the Pearson's correlation coefficient and the root mean-squared error (RMSE; see Table A1). Figure4 and Fig.5 show the time series of total water levels at different tide gauge stations during landfall of TC Irma and ETC Xynthia, respectively. The Pearson's correlation between the GTSM-simulated and observed total water levels is high for both events, indicating a good agreement. For TC Irma, the average correlation across the nine stations is 0.93 with a standard deviation of 0.06 m. For ETC Xynthia, the average correlation across the six stations is 1.00 with a standard deviation of 0.01. Additionally, TC Irma has a RMSE of 0.28 m with a standard deviation of 0.09 m, and ETC Xynthia has a RMSE of 0.22 m with a standard deviation of 0.08 m. This shows that while there are some minor differences between the GTSM simulations and observations, generally there is a good agreement.

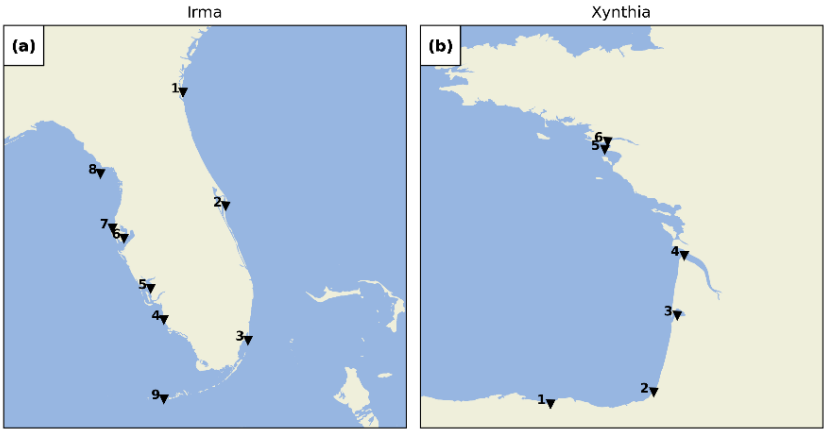

Figure 3. GESLA tide gauge stations for the case studies Irma (panel a) and Xynthia (panel b).

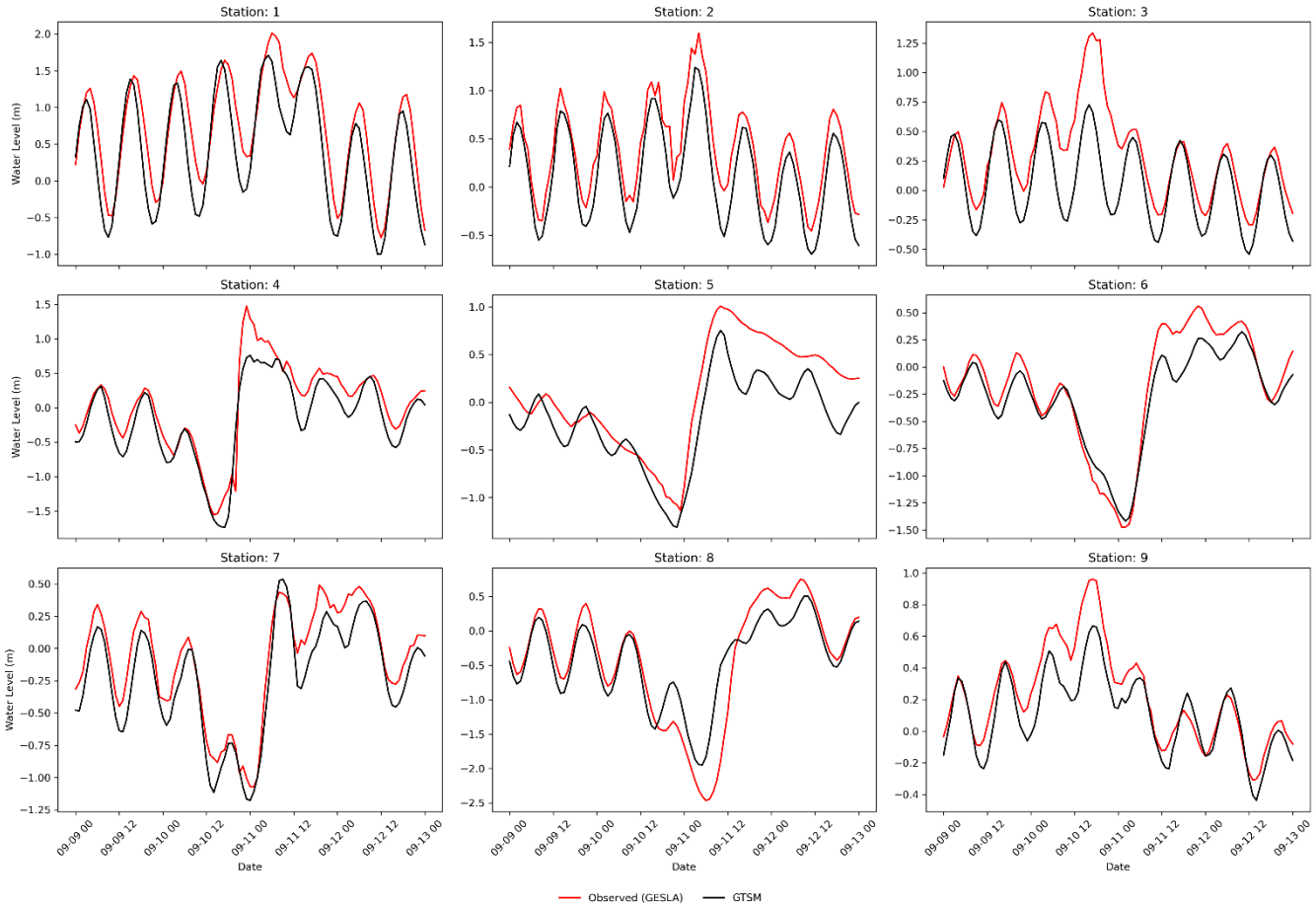

Figure 4. Validation of total water levels for the case study Irma, for the nine tide gauge stations depicted in Fig. 3.

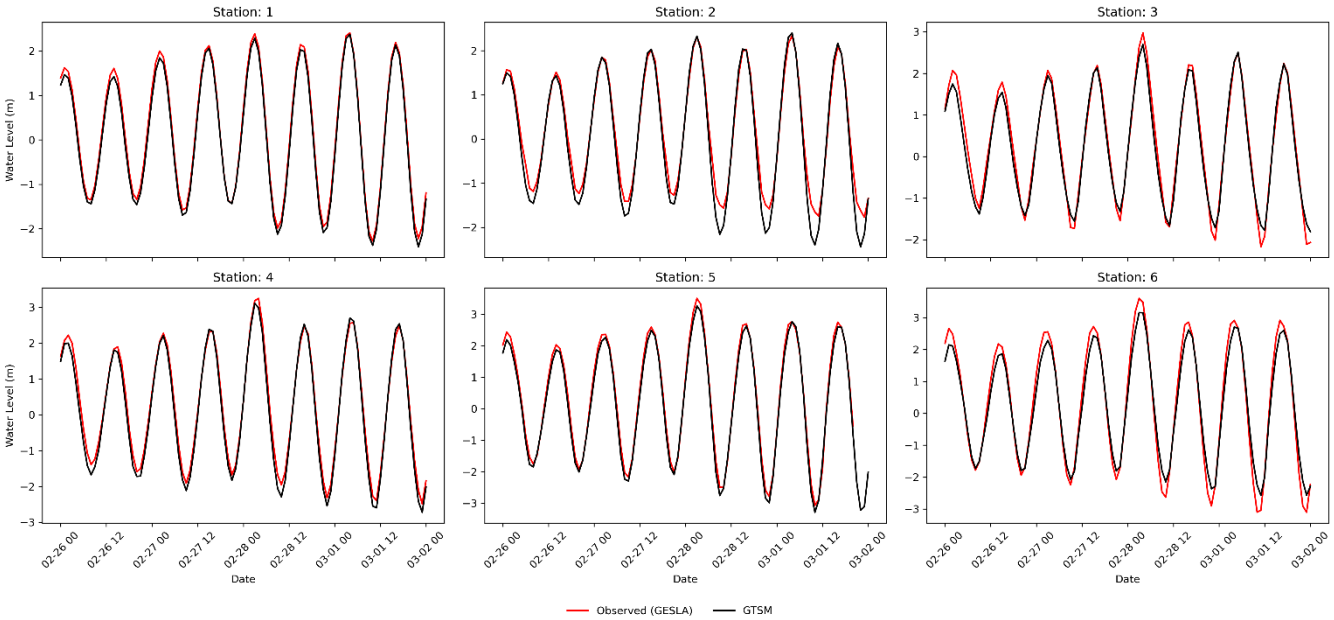

Figure 5. Validation of total water levels for the case study Xynthia, for the six tide gauge stations depicted in Fig. 3."

The validation of each model configuration has also been included in the appendices, and used to interpret the results:

(Lines 232 – 233): "Figure 8 panels a, e and i show the maximum water levels simulated by G1 for three case studies, and depict the maximum observed water levels for various GESLA tide gauge stations."

(Lines 240 – 242): "The small effect of the temporal refinement for TC Irma can be observed as well in Table A1 and Fig. A2, where G1 and G2 present similar timeseries and performance coefficients when compared to observed water levels."

(Lines 251 – 253): "The small effect of the temporal refinement for ETC Xynthia can be observed as well in Table A1 and Fig. A3, where G1 and G2 present similar timeseries and performance coefficients when compared to observed water levels."

(Lines 264 – 268): "Water levels for nine tide gauge stations along the coast indicate that while G1 underestimates the peak of TC Irma in most locations (Fig. A2, all stations but station 7), N1 simulates on average higher peaks, resulting sometimes in overestimations (Fig. A2, station 9). Additionally, the performance of N1 is slightly better than G1 for six tide gauge stations (stations 1-6), as reflected in Table A1, which shows lower RMSE values. However, for stations 7-9, G1 shows slightly higher RMSE and Pearson's correlation."

(Lines 275 – 278): "When comparing the performance of N1 with G1 (Table A1 and Fig. A3), both model configurations can predict the timeseries pattern well, with high Pearson's correlation coefficients. Overall, the RMSE for Xynthia is similar for most tide gauge stations, except for two stations located in the mouth of estuaries (stations 3 and 6)".

(Lines 281 – 285): "For TC Irma (Fig. 8 panel c), the nesting of a local model at high-resolution with updated GEBCO2023 bathymetry results in maximum water levels that are 0.3 m higher than G1 in the south of Florida. Compared to N1, model configuration N2 provides slightly higher water levels south of Florida. Those differences come from differences between GEBCO2023 and GEBCO2019 in the region. N2 shows a similar performance to G1 and N1 across nine tide gauge stations (Table A1 and Fig. A2)."

(Lines 288 – 294): "For ETC Xynthia (Fig. 8 panels l), the water levels from the nested local model at high-resolution with GEBCO2023 are lower overall than water levels for G1. Compared to N1, the model configuration N2 provides a similar pattern of water level decrease, however, the maximum water level reduction compared to G1 is slightly less than for N1. The performance of N2, as shown in Table A1 and Fig. A3, is comparable to that of G1 and N2, except at two tide gauge stations (station 3 and 6) where GEBCO2023 does not accurately capture the bathymetry of the river channels in the estuaries. In contrast, EMODNET2018, the bathymetry used in model configurations N1 and N3, better resolves these details (see Fig. A7)."

Table A1. Validation indicators that compare the maximum total water levels and observations of GESLA for the case studies Irma and Xynthia.

Irma	RMSE [m]				Pearson correlation [-]			
Station	**G1**	**G2**	**N1**	**N2**	**G1**	**G2**	**N1**	**N2**
1	0.41	0.41	0.39	0.40	0.92	0.92	0.92	0.92
2	0.28	0.27	0.25	0.25	0.98	0.98	0.98	0.98
3	0.33	0.33	0.32	0.33	0.79	0.78	0.81	0.79
4	0.27	0.26	0.21	0.24	0.96	0.96	0.96	0.94
5	0.35	0.35	0.33	0.31	0.93	0.93	0.93	0.93
6	0.18	0.18	0.17	0.21	0.98	0.98	0.98	0.94
7	0.17	0.17	0.14	0.14	0.97	0.97	0.95	0.95
8	0.39	0.39	0.42	0.45	0.92	0.92	0.90	0.88
9	0.16	0.16	0.18	0.10	0.93	0.92	0.90	0.96
Average	***0.28***	***0.28***	***0.27***	***0.27***	***0.93***	***0.93***	***0.93***	***0.92***
Standard deviation	***0.09***	***0.09***	***0.10***	***0.11***	***0.06***	***0.06***	***0.05***	***0.05***
Xynthia	**RMSE [m]**				**Pearson correlation [-]**			
Station	**G1**	**G2**	**N1**	**N2**	**G1**	**G2**	**N1**	**N2**
1	0.12	0.13	0.13	0.13	1.00	1.00	1.00	1.00
2	0.27	0.29	0.22	0.26	0.99	0.99	0.99	0.99
3	0.21	0.20	0.47	0.61	0.99	0.99	0.95	0.91
4	0.20	0.21	0.19	0.34	1.00	1.00	1.00	0.98
5	0.18	0.18	0.24	0.25	1.00	1.00	0.99	0.99
6	0.34	0.31	0.49	0.92	0.99	0.99	0.98	0.90
Average	***0.22***	***0.22***	***0.29***	***0.42***	***1.00***	***1.00***	***0.99***	***0.96***
Standard deviation	***0.08***	***0.07***	***0.15***	***0.29***	***0.01***	***0.01***	***0.02***	***0.04***

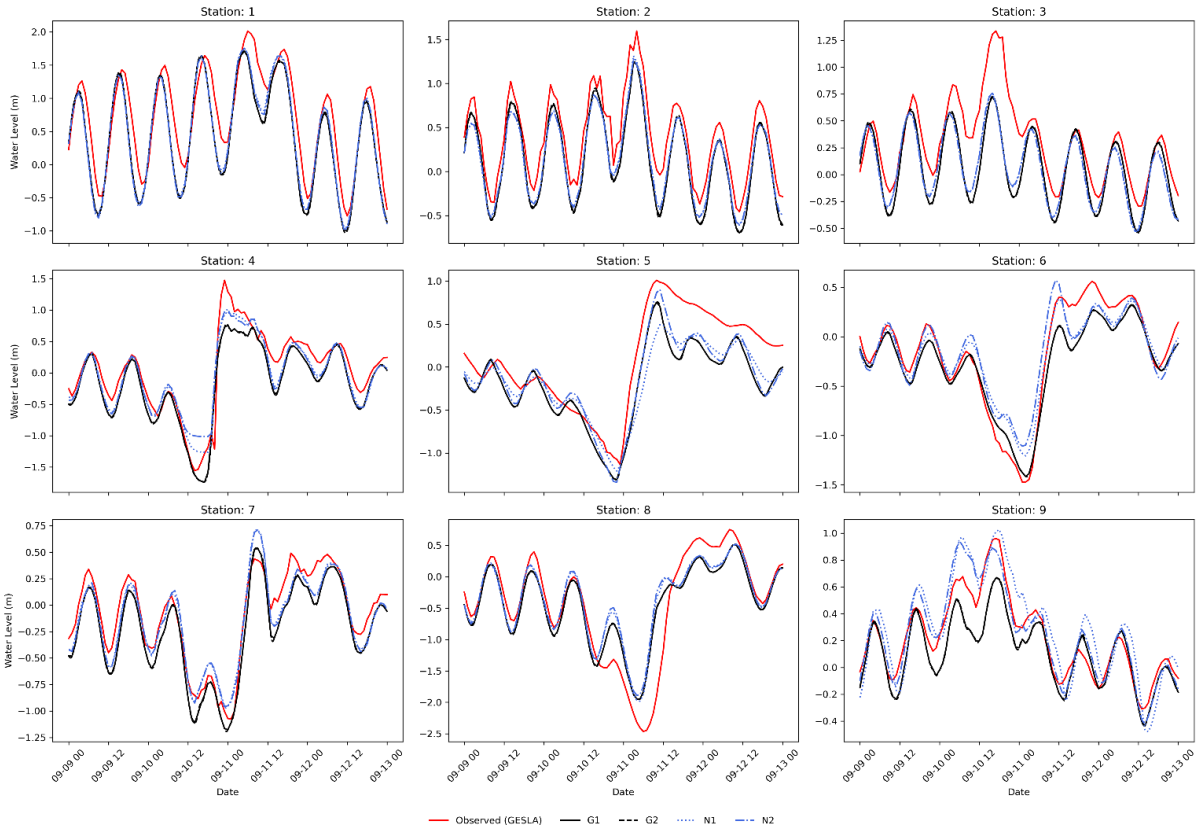

Figure A2. Validation of total water levels for the case study Irma, for the nine locations depicted in Fig. 3.

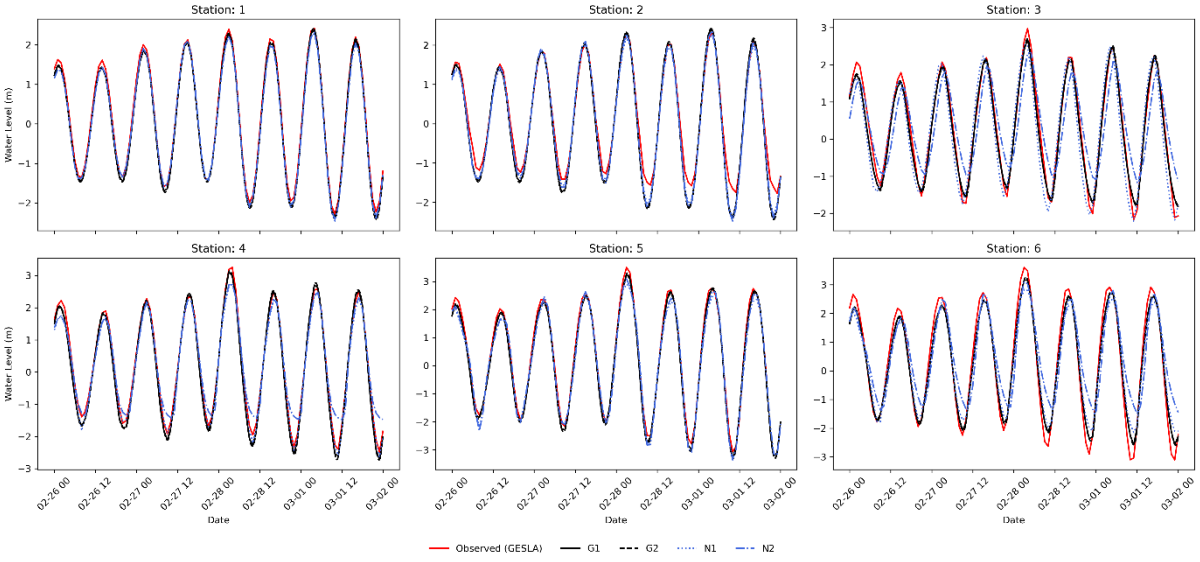

Figure A3. Validation of total water levels for the case study Xynthia, for the six locations depicted in Fig. 3.

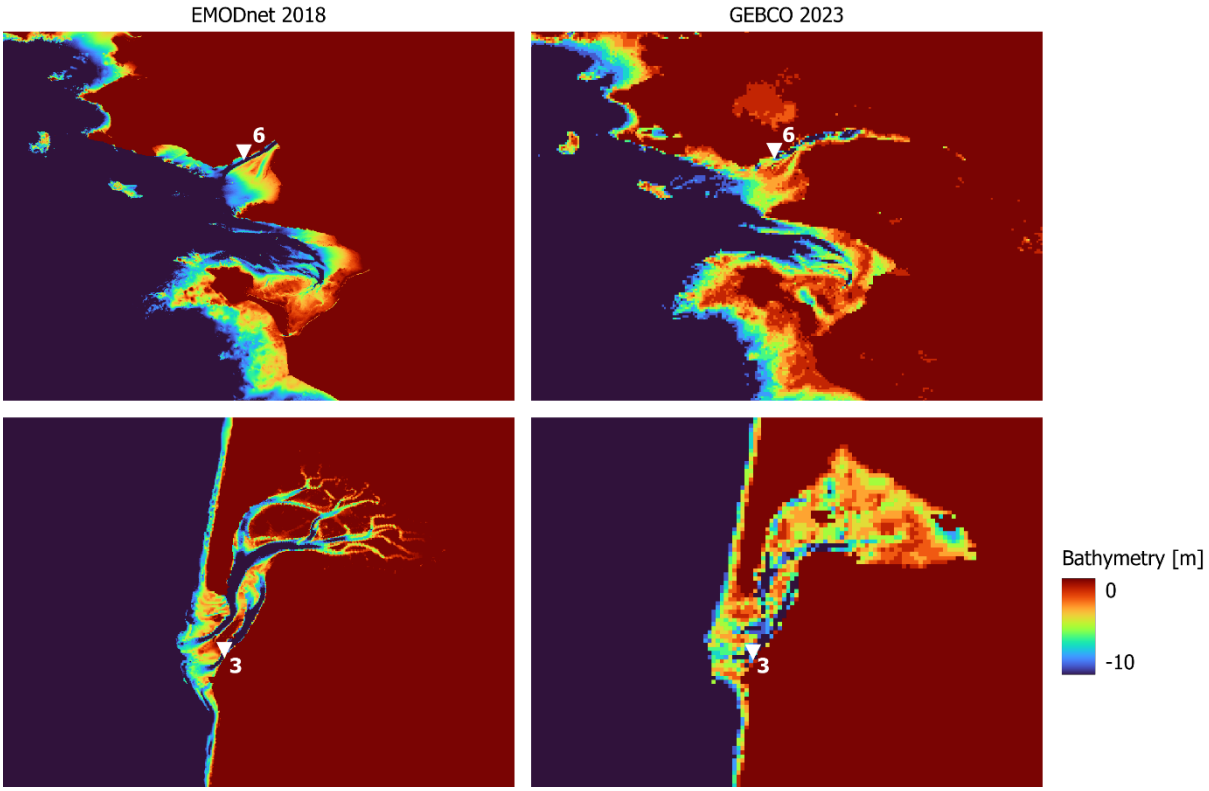

Figure A7. Close look at the bathymetry of two stations (top row: station 3 and bottom row station: 6) that provide lower performance with updated bathymetry, for the case study Xynthia. Left: Bathymetric map of EMODNet2018. Right: Bathymetric map of GEBCO2023.

We validated the flood extents of the modelling framework for case study Xynthia, calculating the hit rate, false-alarm ratio and critical success index:

(Lines 183 – 198): "To validate the hydrodynamic flood hazard modelling component of the modelling framework, we compare the modelled flood extents with observed flood extents derived from field measurements. This comparison is done for Xynthia, the only case study for which observed flood extent data are available (Breilh et al., 2013; DDTM, 2011). We measure the model skill using: (1) the hit rate (H), defined as the flood area correctly simulated over the observed flooded area (Eq (1)); (2) the false-alarm ratio (F), defined as the area wrongly simulated over the observed flooded area (Eq (2)); and (3) the critical success index (C), defined as the area correctly simulated to be flooded over the union of the observed and modelled flooded area (Eq (3)). Figure 6 shows the skill of the modelled maximum flood extents by SFINCS using the GTSM water levels as forcing. The hit rate is 0.78, correctly representing the flooding in most regions, only underestimating it in regions further inland. The false-alarm ratio of the model is 0.62. Flooding is overestimated in the north, likely due to the lack of flood protection measures included in the model that are present in reality. The critical success index is 0.48, as a result of the areas well simulated and those over and underpredicted. While the performance of the flood model is negatively affected by the quality of the topography and the representation of local features such as dikes, we consider the performance sufficient for large-scale modelling and comparable to other studies such as (Ramirez et al., 2016; Vousdoukas et al., 2016b).

$$H = \frac{F_{modelled} \cap F_{observed}}{F_{observed}} \quad (1)$$

$$F = \frac{F_{modelled} / F_{observed}}{F_{observed}} \quad (2)$$

$$C = \frac{F_{modelled} \cap F_{observed}}{F_{modelled} \cup F_{observed}} \quad (3)$$

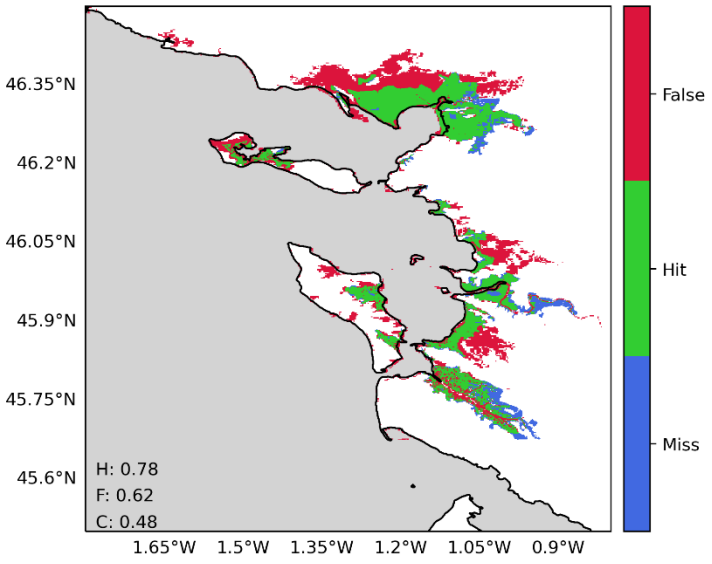

Figure 6. Validation of the flood hazard modelling component of the modelling framework for the case study Xynthia, using the water levels of the default configuration of GTSM as a forcing. The maps compare the modelled and observed maximum flood extents, where: green indicates flood areas correctly simulated; blue flood areas not simulated but observed; and red flood areas simulated but not predicted. Performance indicators for the hit rate (H), false-alarm ratio (F) and critical success index (C) are shown in each panel."

We have also used the validation of each model configuration to interpret the results for the flood hazard modelling part:

(Lines 322 – 323): "For ETC Xynthia, G2 shows a higher hit rate and false-alarm ratio compared to G1, but the same critical success index (see Fig. A9)."

(Lines 330 – 331): "For ETC Xynthia, G3 shows the same hit rate as G1, higher false-alarm ratio and the same critical success index (see Fig. A9)."

(Lines 339 – 340): "For ETC Xynthia, N1 shows a lower hit rate and false-alarm ratio compared to G1, and the same critical success index (see Fig. A9)."

(Lines 350 – 351): "For ETC Xynthia, N2 shows a lower hit rate and false-alarm ratio compared to G1, and the same critical success index (see Fig. A9)."

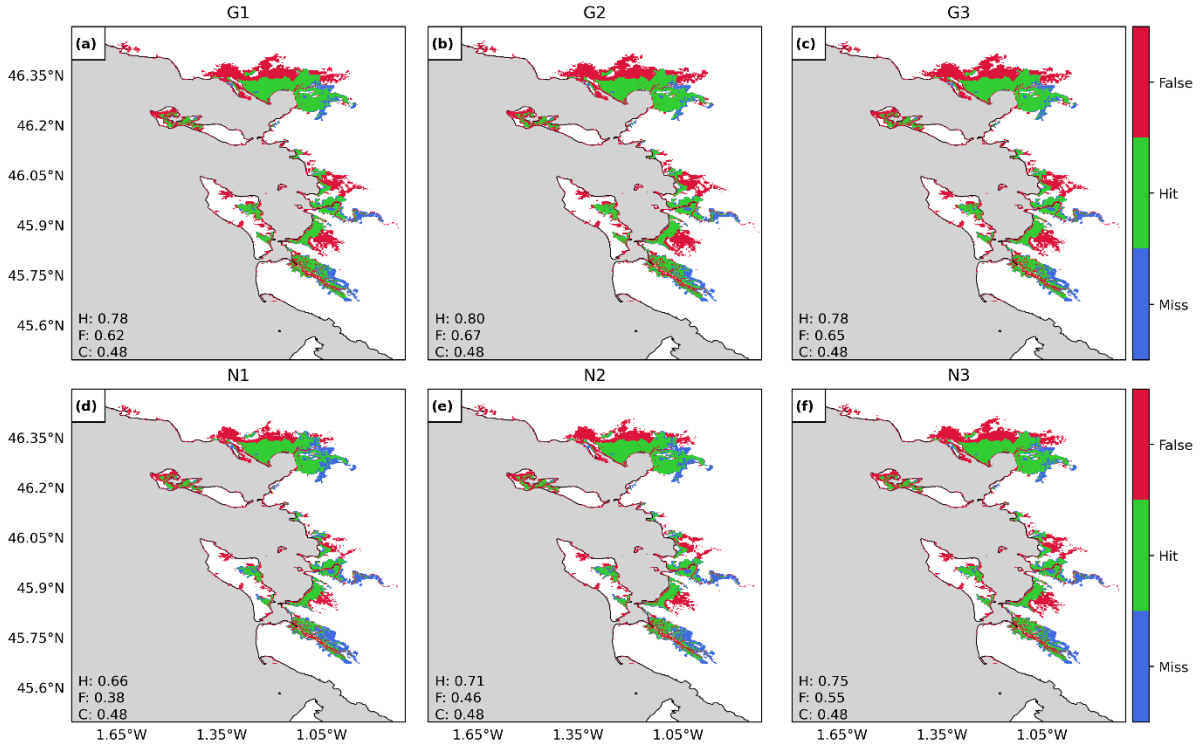

Figure A9. Validation of flood extents for the case study Xynthia against observed flood extents. The maps compare the modelled and observed maximum flood extents for each model configuration, see Table 1, where: green indicates flood areas correctly simulated; blue flood areas not simulated but observed; and red flood areas simulated but not predicted. Performance indicators for the hit rate (H), false-alarm ratio (F) and critical success index (C) for each configuration are shown in each panel.

The model validation does not provide a conclusive result in terms of which model configuration provides the best results. This is really case dependent, and it is difficult to validate a global model with the limited tide gauge stations available. Nevertheless, no model configuration shows bad results, and the flexibility and easiness of MOSAIC can be used by users to explore which settings are best for their specific case study and region of interest. We have included these conclusions in the manuscript as follows:

> *(Lines 401 – 432): "The results of the sensitivity analysis conducted in this study reveal the complexity of hydrodynamic modelling and the sensitivity to specific local settings and storm characteristics. A comparison of the fully refined N3 configuration with the default G1 configuration reveals differing behaviours across the case studies in terms of changes in water levels and flood depths, both spatially and in magnitude. For instance, model configuration N3 simulates higher water levels almost everywhere for TC Irma. However, for TC Haiyan and ETC Xynthia, certain regions show higher water levels with N3, while others show lower water levels compared to G1. Similarly, flood depths around Jacksonville for TC Irma are generally higher with N3, although some areas experience lower values. In contrast, for TC Haiyan in Tacloban, flooding significantly increases with N3, whereas for ETC Xynthia flood depths decrease notably around La Rochelle.*
>
> *Refining the temporal output resolution (model configuration G2) has a significant influence on small, rapidly intensifying TCs like Haiyan, resulting in water levels and flood depths that are 2 m and 1 m higher, respectively,*

compared to G1. However, for ETCs, the refinement of temporal output resolution does not lead to substantial changes in water levels or flood depths, with a 1-hourly temporal resolution providing sufficiently accurate results. Refining the spatial output locations of GTSM (model configuration G3) provides more coastal boundary conditions for SFINCS. For regions where the water levels have more spatial heterogeneity along the coast, this refinement becomes most relevant. For TC Haiyan, for example, the coastal boundary conditions in the bay of Tacloban raised from 4 locations to more than 20 (see Fig. 7), leading to flood depths 1 m higher than G1. Furthermore, regions with more complex topographies such as the south of Florida for TC Irma or the Tacloban bay for TC Haiyan are influenced by the grid refinement of N1, leading to larger differences with G1 in terms or water levels and consequently, flooding. The choice of bathymetric datasets also plays a role in the prediction of water levels, contributing to the differences observed between N1 and N2 in all the case studies. Based on these results, we can conclude that the refinement of the global modelling approach can significantly impact the simulation of coastal water levels and flood depths at local scale, although the differences in local settings make that there is no one-size-fits-all approach.

The validation of the model configurations for the different case studies also highlights the complexities involved in refining hydrodynamic models, and how each specific setting impacts overall performance. It is challenging to assess the performance of global models due to the limited number of tide gauge stations available, meaning the validation results might not be fully representative over the entire domain. Another source of uncertainty is the location of these tide gauge stations, which are often situated in enclosed basins or harbours, where hydrodynamic models have more difficulty simulating water levels compared to open sea conditions. While the performance indicators from this study, with Pearson's correlations above 0.92 and RMSEs in general less than 0.3 m, suggest that all the refinements perform adequately and similarly to G1, the validation does not allow us to determine which model configuration consistently provides the best overall performance. This outcome largely depends on the storm characteristics and regional topography. However, the flexibility and ease of use of MOSAIC, as a Python-based framework, make it a valuable resource for users to further explore which are the optimal settings for their case study and region of interest."

The effects of waves (especially infragravity energy), precipitation, and river discharge should be addressed and discussed more thoroughly in the introduction (other methodologies that take into account these factors) and in the discussion (what effects could be missed) .

We have included a reflection on the flood drivers that are considered in this manuscript and the ones that could be applied later on within MOSAIC:

(Lines 62 – 70): "Additionally, large-scale hazard assessments typically focus on a single flood driver (Tiggeloven et al., 2020; Vousdoukas et al., 2018b; Ward et al., 2020). However, TC and ETC events often produce precipitation, river discharge, storm surges and waves, all of which can contribute to flooding. When these drivers occur in combinations, they can significantly amplify flood hazards and risks. For instance, recent research showed that storm surge exacerbates fluvial flooding at global scale (Eilander et al., 2020). Few studies have analysed the effects and interactions of multiple flood drivers. While Bates et al. (2021) performed a combined risk assessment of fluvial, pluvial and coastal flooding for the continental USA, Eilander et al. (2023) introduced the first globally-applicable compound flood modelling framework that accounts for

precipitation, river discharge and storm tides. However, the inclusion of waves in large-scale assessments and the interactions between flood drivers remains a challenge."

(Lines 456 – 462): "In this study, we have implemented MOSAIC to simulate coastal flooding driven by storm surges. However, since flooding typically results from a combination of various drivers, our results currently underestimate flooding near estuaries and deltas due to the exclusion of precipitation and river discharge, and near steep coasts due to the exclusion of waves and overtopping. Future research on TCs and ETCs may further develop MOSAIC and include other drivers such as waves, rainfall and discharge. Considering that HydroMT and SFINCS are capable of handling compound flooding induced by pluvial and fluvial drivers (Eilander et al., 2023), there is potential for future enhancements of MOSAIC to incorporate the modelling of compound events."

Specific question:

Local high resolution model domain: why these domains have been selected? Have they been defined based on the cyclone tracks?

Yes, these domains were selected based on the tropical cyclone tracks, and also considering that the size of the model domain was not too small, given that that could cause model instabilities.

Minor comments:

GEBCO 2020 in line 119 while GEBCO2023 is mentioned in Table 1.

GEBCO2020 is used for the flood hazard modelling using SFINCS, while the updated bathymetry of GTSM is GEBCO2023 (this is what Table 1 refers to).

Realistic configuration in Figure 8: not sure what result/output is this.

This was a typo. We have updated the figure with the new nomenclatures:

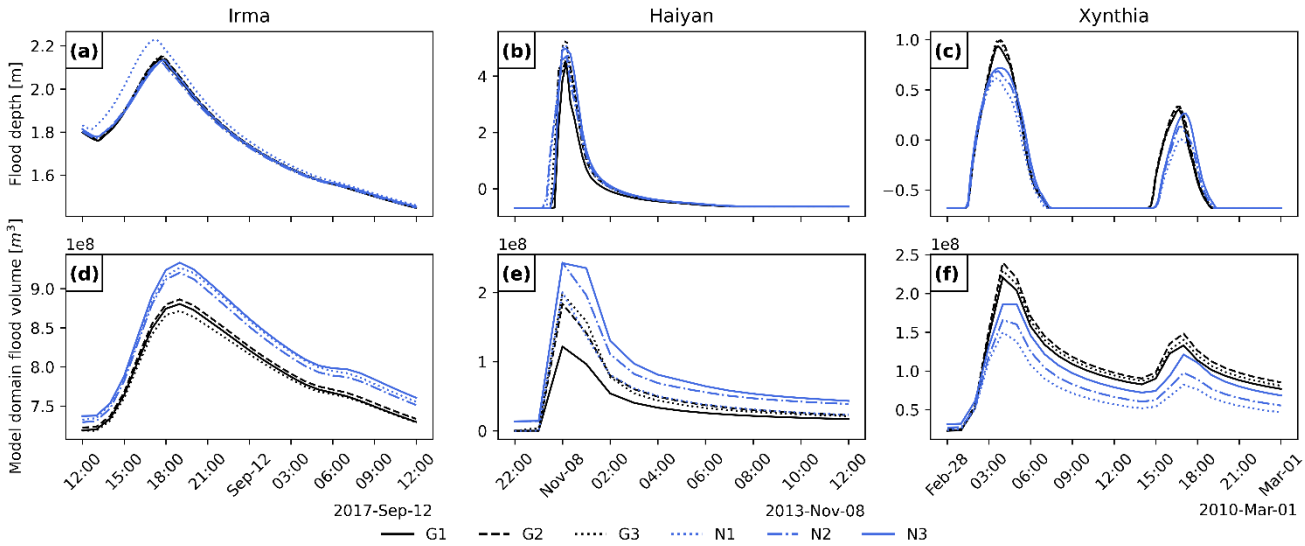

Figure 12. Flood depth timeseries for three observation points and flood volume timeseries for the SFINCS model domain of each case study. The spatial location of the SFINCS output point locations can be observed in Fig. 11 panels a, d, g.