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

Responses to reviewers:

Reviewer #1:

Dear Authors,

After reviewing your manuscript, I find that it holds significant relevance and potential. However, in its current form, it fails to address any substantial scientific question. The results presented are merely a model-model comparison, illustrating that variations in spatial or temporal resolution affect the outcomes. This does not address the scientific questions or limitations introduced in the introduction. Therefore, I recommend major revisions and provide the following comments to help strengthen the manuscript and enhance its scientific rigor.

We are pleased to read that the reviewer finds our paper has potential to provide a valuable scientific contribution. In response to the detailed and constructive comments, we have significantly revised the paper.

We have rewritten the introduction to clearly outline the limitations addressed by the manuscript. In the manuscript we now present a modelling framework that is easy to use and flexible, and 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.

We believe this has greatly improved the manuscript and we thank the reviewer for the time taken to review our manuscript.

1. **Introduction**: This section is well-written but could benefit from more specifics regarding the dynamic processes that are currently missing (L46-51). For instance, details on wave-driven processes, hydrological processes, and man-made structures would be valuable. Additionally, I challenge the notion that the limitation of topo-bathymetry in global applications can be resolved solely through grid refinement. In my mind, there are three main methodological challenges: resolution, input data sets (topo-bathy and others), and physical processes. This paper addresses the first one but not the other two. Hence, the linkage from the scientific gap to the approach does not hold, as the MOSAIC modeling framework does not resolve challenges with input data nor does it address additional processes relevant for inundation that global models fail to account for.

Thank you for the insightful suggestions regarding the introduction. Following your recommendation, we have restructured the main methodological challenges into three sections. Additionally, we have added the inclusion of multiple flood drivers as a challenge for flood risk assessments. Finally, the *manuscript has been revised to explicitly state that our primary aim is to benchmark the implications of model resolution:*

> *(Lines 48 – 84): "We identify here three main methodological limitations of large-scale hazard assessments. First, coastal geometry strongly influences extreme sea levels (Bloemendaal et al., 2019; Dullaart et al., 2020; Mori et al., 2014; Woodruff et al., 2023), with large variability at local scale. Consequently, in regions with complex morphologies, such as estuaries, semi-enclosed bays or barrier systems, global models lack the resolution required to accurately resolve the extreme sea levels (Bunya et al., 2010; Dietrich et al., 2010). Grid refinement and nesting of local high-resolution models within coarser global models can result in improved coastal boundary conditions. Pelupessy et al. (2017) used a similar multiscale approach to obtain realistic boundary conditions by nesting a global circulation model and a high-resolution barotropic model. Second, the accuracy of input datasets such as the meteorological forcing and the bathymetry have large influence on the total water levels. Coarse meteorological forcings – both in terms of spatial and temporal resolution – might not be able to capture the resolution necessary to resolve intense storms, while errors in the bathymetric datasets will propagate to the modelling of storm surge levels (Bloemendaal et al., 2019; Dullaart et al., 2020). Third, coastal flooding is a dynamic process where flood duration and physical processes play a key role. However, given the high computational costs associated with using hydrodynamic flood models, their use has been limited to local application. Most large-scale hazard assessments have used static flood modelling methods, which neglect flood dynamics (Hinkel et al., 2014; Muis et al., 2016; Ramirez et al., 2016; Vafeidis et al., 2019; Vousdoukas et al., 2016b). 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.*
>
> *In this study, we introduce the MOSAIC (MOdelling Sea Level And Inundation for Cyclones) modelling framework with the aim of providing a flexible Python-based modelling framework that allows to dynamically simulate TC and ETC water levels and coastal flooding events. To analyse the effects of model resolution, MOSAIC applies a multiscale modelling approach in which local models with high-resolution (~45 m to 25 km) are nested within a large-scale model with a coarser resolution (~2.5 km to 25 km). To enable hydrodynamic flood modelling, MOSAIC couples two existing modelling approaches: (1) to simulate water levels generated from storm surges and tides at global to local scale it couples the hydrodynamic Global Tide and Surge Model (GTSM) and Delft3D Flexible Mesh software; and (2) to dynamically simulate overland flooding at local scale it couples the simulated water levels with the Super-Fast INunadation of CoastS model (SFINCS). We use a*

reproducible approach that is globally applicable and that can automatically generate local Delft3D Flexible Mesh models as well as local SFINCS models. In this study, we showcase the potential of the MOSAIC framework by applying it to three case studies where large storm surges caused catastrophic flooding events, namely historical storm events TC Irma, TC Haiyan, and ETC Xynthia (see Fig. 1; Bertin et al., 2012; Cangialosi et al., 2018; Lapidez et al., 2015). For each of these storms, we simulate the coastal water levels and flood depths. Moreover, we perform a sensitivity analysis of different modelling settings with the goal of benchmarking model configurations with different resolutions."

2. **MOSAIC Modeling Framework**: The authors aim to introduce a modeling framework. To do so successfully, a more comprehensive introduction to other modeling frameworks and/or nesting techniques is necessary. Additionally, more details are needed on what has been specifically programmed and what is novel about it. For example, details on the Holland parametric wind model and how it is integrated are missing. I also miss details on the nesting procedure used for both the offline Delft3D FM and SFINCS approach. A more rigorous description of the code would enhance the scientific value of the manuscript.

Thank you for your valuable suggestions. In response to your comments, we have included more information about current modelling frameworks. Specifically, we now present a modelling framework that uses model nesting to improve the outputs in ocean models, and extends to land to hydrodynamically simulate compound flooding. The manuscript has been updated accordingly as follows:

(Lines 52 – 54): "Grid refinement and nesting of local high-resolution models within coarser global models can result in improved coastal boundary conditions. Pelupessy et al. (2017) used a similar multiscale approach to obtain realistic boundary conditions by nesting a global circulation model and a high-resolution barotropic model."

(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."

Using the Holland model is an option within the modelling framework that can be used when the user considers it necessary. To enhance the clarity on the Holland parametric wind model, we have added the following lines to the manuscript, where we define more details of the Holland model and how it is used in the framework:

> (Lines 103 – 113): “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. The tropical cyclone track data is retrieved from the Joint Typhoon Warning Center at 6 hourly intervals (Naval Meteorology and Oceanography Command, 2022) and is converted to a polar grid with 36 radial bins, 375 arcs and a radius of 350 km using the Holland parametric wind model (Holland et al., 2010). Following the methodology of Dullaart et al. (2021) and Lin and Chavas (2012), we apply a counter-clockwise rotation angle of $\beta = 20^\circ$ and set the storm translation to surface background wind reduction factor at $\alpha = 0.55$. Additionally, we use an empirical surface wind reduction factor (SWRF) of 0.85 (Batts et al., 1980), and convert 1-minute average winds to 10-minute averages using a factor of 0.915 (Harper et al., 2010). The Holland model’s output provides a file that defines a polar grid containing pressure and wind fields. To extend the pressure and wind fields beyond the Holland model’s defined TC boundary, we linearly interpolate these fields on the outermost 75% to align with the ERA5 background data (Deltares, 2024).”

In order to improve the explanation of the nesting procedure used for nesting GTSM and the local Delft3D FM model, we have updated the manuscript as follows:

> (Lines 152 – 157): “Second, MOSAIC uses an offline coupling approach to nest the local Delft3D Flexible Mesh model within GTSM. A Python script is used to first identify the boundaries of the local Delft3D Flexible Mesh model. These boundaries are then used to determine the specific locations where GTSM output should be extracted. Subsequently, GTSM provides the water level timeseries at the boundaries of the local model. Finally, the local high-resolution model is executed using the water levels derived from GTSM as forcing input, together with the same meteorological forcing as for GTSM.”

Although we previously mentioned using HydroMT to couple GTSM and SFINCS, the specific output from GTSM that serves as input for SFINCS was not clearly explained. To address this, we have updated the manuscript with the following clarifications:

> (Lines 125 – 126): “GTSM provides as output water level timeseries over a grid in the ocean and for locations along every ~5 km of the coast.”

And:

> (Lines 175 – 182): “To build the SFINCS models and couple them with GTSM, MOSAIC uses the HydroMTv0.7.1 (Hydro Model Tools) package (Eilander et al., 2023). HydroMT is an open-source Python package, which provides automated and reproducible model building and analysis of results. HydroMT uses a modular approach in which datasets and model setup configurations can easily be interchanged. In the MOSAIC framework presented in this paper, we take advantage of HydroMT in several ways: (1) to automatically convert the forcing files from GTSM and the other input into the model specific input format; (2) to easily build a reproducible SFINCS model; and (3) to perform the analysis of the SFINCS model output. SFINCS is forced with GTSM water level timeseries at locations along every ~5 km of the coastline, and provides as output water level timeseries for each grid cell. Finally, flood depth maps are derived from the maximum water levels by subtracting the DEM elevation.”

3. **Modeling Results**: This section requires the most work. As mentioned, in its current state, it is a model-model comparison without any significant insights. To address this, I strongly recommend the authors include relevant observations of observed water levels and flood extents. Without these, it is difficult to assess differences and model accuracy.

Thank you for the suggestions. We understand the concern about including relevant observations of water levels and flood extents. For this reason, we have thoroughly revised this section and included a validation of the results of the case studies for which we had observations. We also improved the sensitivity analysis and provided additional results of this analysis in the appendix. Overall, we believe the result section is much stronger now. Note that 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 elaborate on the changes made in the manuscript below.

We have validated the total water levels using observations from the GESLA tide gauge stations for the case studies Irma and Xynthia (Haigh et al., 2023). 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 also 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 at the background. We updated the manuscript as follows:

> *(Lines 101 – 113): "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. The tropical cyclone track data is retrieved from the Joint Typhoon Warning Center at 6 hourly intervals (Naval Meteorology and Oceanography Command, 2022) and is converted to a polar grid with 36 radial bins, 375 arcs and a radius of 350 km using the Holland parametric wind model (Holland et al., 2010). Following the methodology of Dullaart et al. (2021) and Lin and Chavas (2012), we apply a counter-clockwise rotation angle of β = 20° and set the storm translation to surface background wind reduction factor at α = 0.55. Additionally, we use an empirical surface wind reduction factor (SWRF) of 0.85 (Batts et al., 1980), and convert 1-minute average winds to 10-minute averages using a factor of 0.915 (Harper et al., 2010). The Holland model's output provides a file that defines a polar grid containing pressure and wind fields. To extend the pressure and wind fields beyond the Holland model's defined TC boundary, we linearly interpolate these fields on the outermost 75% to align with the ERA5 background data (Deltares, 2024)".*

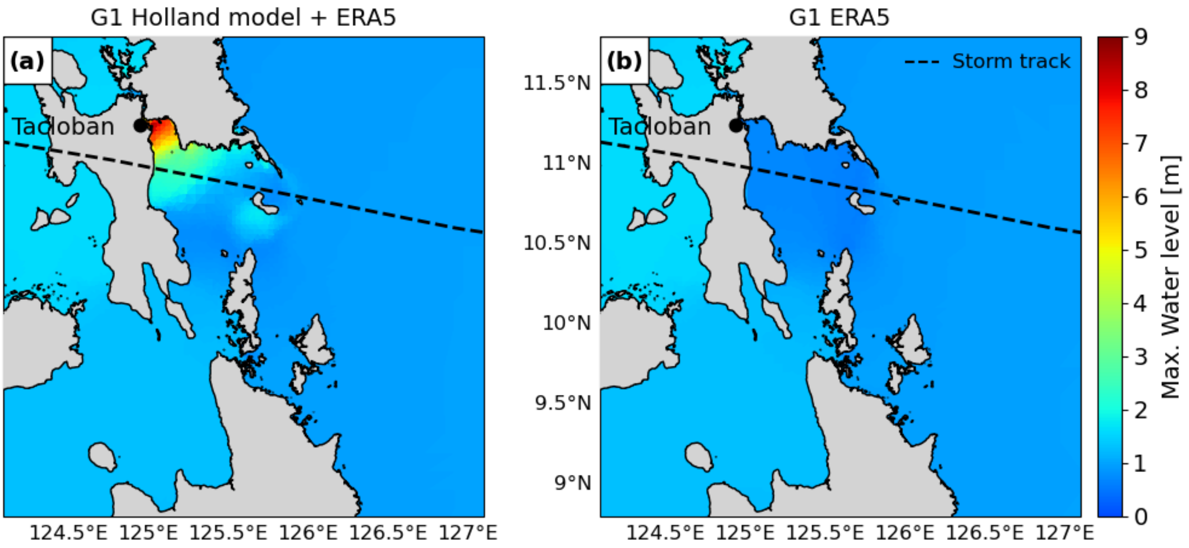

Figure A1. Maximum water levels output of GTSM, for case study Haiyan, with different meteorological forcings. Left: maximum total water levels with ERA5 as forcing. Right: Maximum total water levels with the Holland model combined with ERA5 as a forcing.

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 (Section2.1.1; 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.

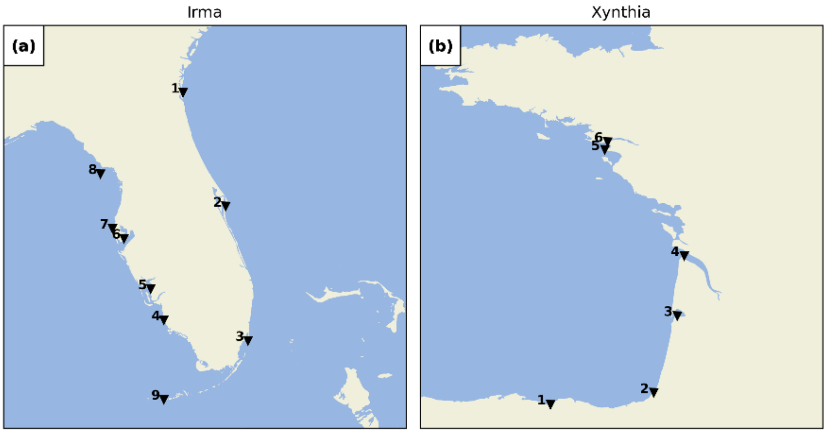

Figure 3. GESLA tide gauge stations for the case studies Irma (panel a) and Xynthia (panel b).

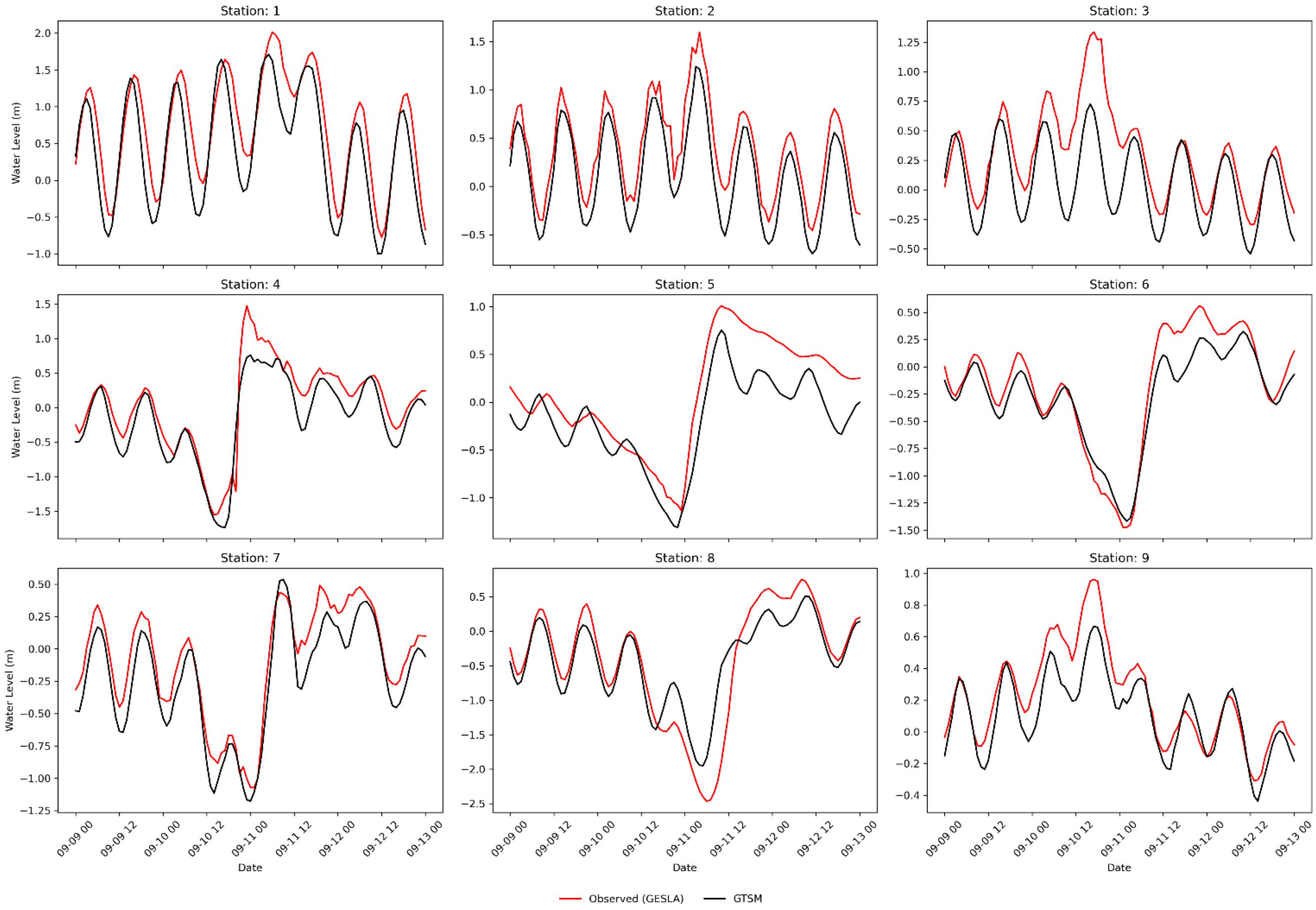

Figure 4. Validation of total water levels for the case study Irma, for the nine tide gauge stations depicted in Fig. 3.

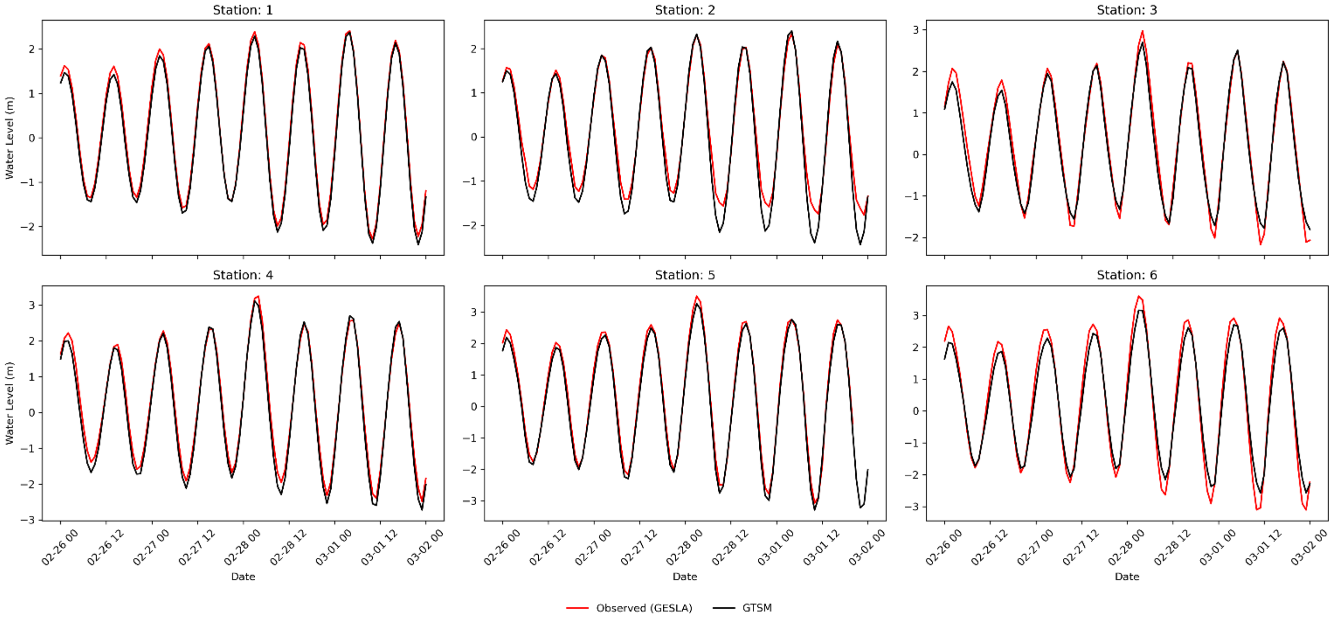

Figure 5. Validation of total water levels for the case study Xynthia, for the six tide gauge stations depicted in Fig. 3."

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***

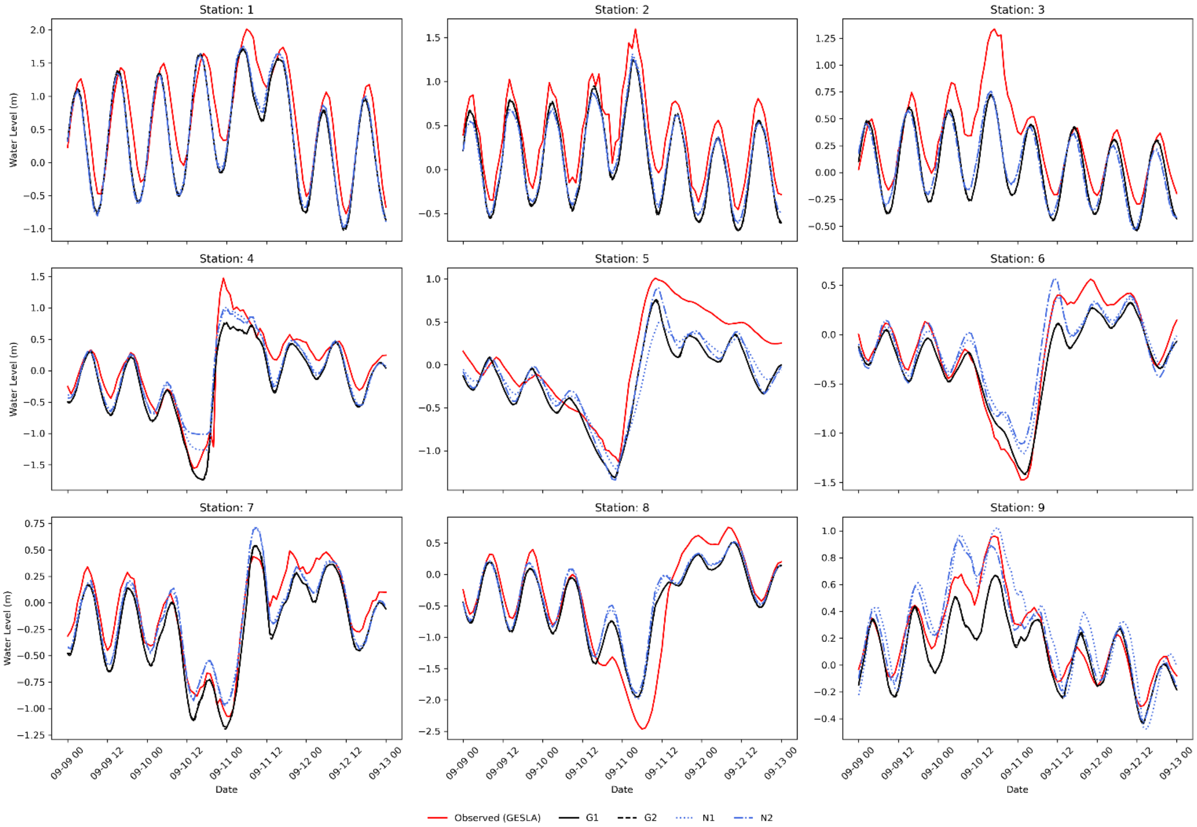

Figure A2. Validation of total water levels for the case study Irma, for the nine locations depicted in Fig. 3.

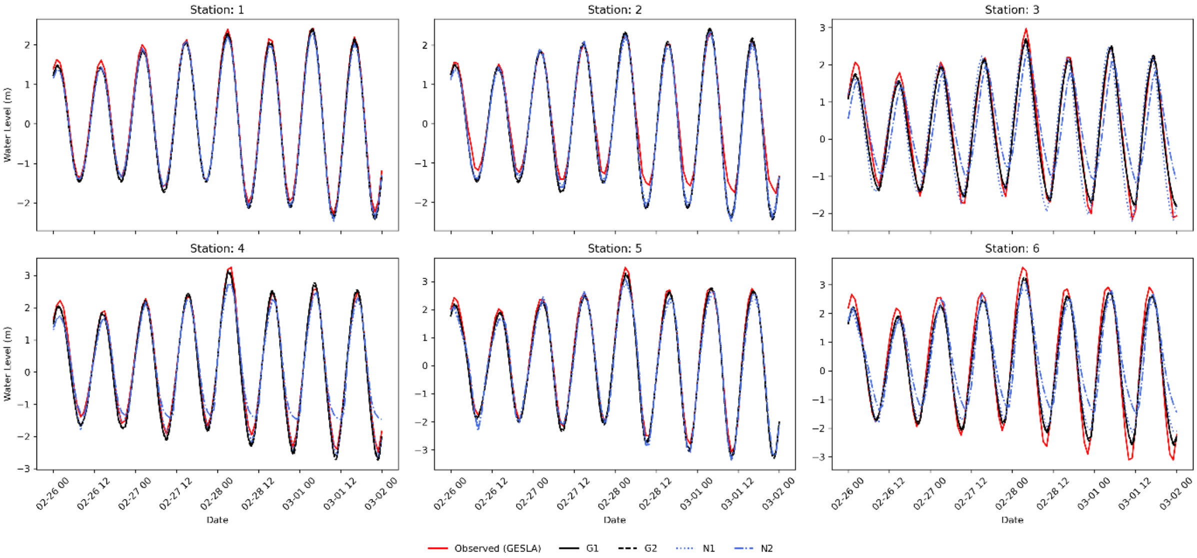

Figure A3. Validation of total water levels for the case study Xynthia, for the six locations depicted in Fig. 3.

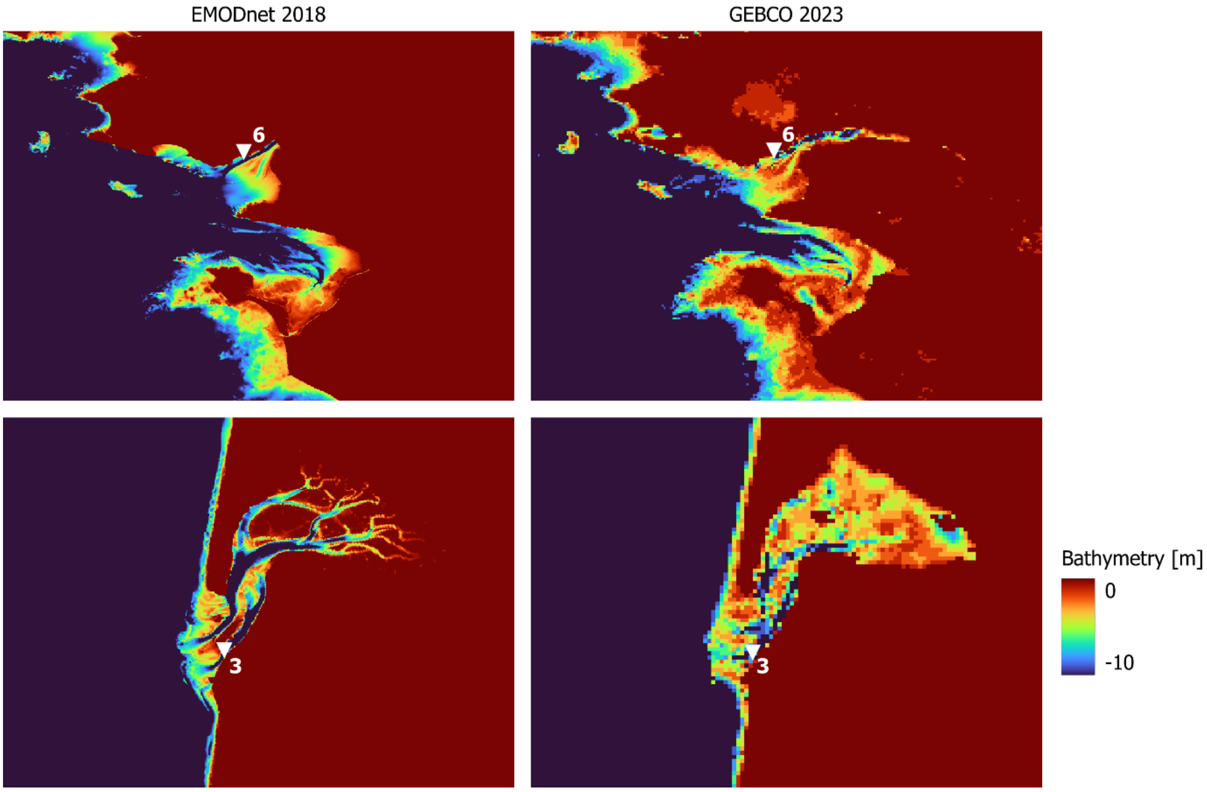

Figure A7. Close look at the bathymetry of two stations (top row: station 3 and bottom row station: 6) that provide lower performance with updated bathymetry, for the case study Xynthia. Left: Bathymetric map of EMODNet2018. Right: Bathymetric map of GEBCO2023.

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

> *(Lines 183 – 195): "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).*

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

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

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

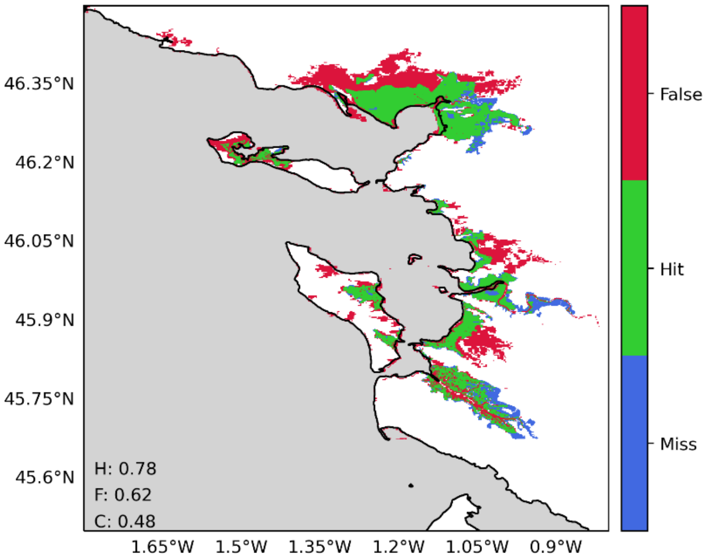

Figure 6. Validation of the flood hazard modelling component of the modelling framework for the case study Xynthia, using the water levels of the default configuration of GTSM as a forcing. The maps compare the modelled and observed maximum flood extents, where: green indicates flood areas correctly 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)."*

This shows that for ETC Xynthia the G@ models can correctly represent the flooded areas due to higher hit rates, but overestimate the flooding more in the northern regions, while the N@ models underestimate the flooding in the south due to the lower water levels simulated in that region.

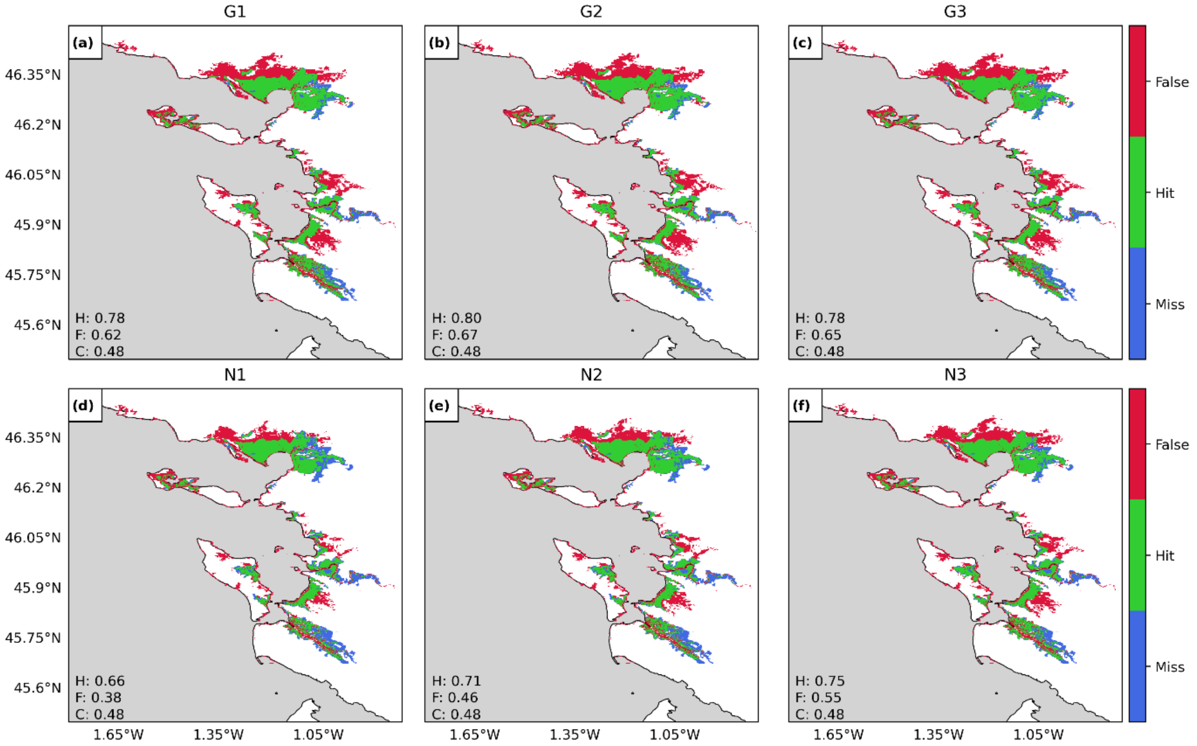

Figure A9. Validation of flood extents for the case study Xynthia against observed flood extents. The maps compare the modelled and observed maximum flood extents for each model configuration, see Table 1, where: green indicates flood areas correctly simulated; blue flood areas not simulated but observed; 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 insights regarding the relevance of temporal and spatial resolution require more simulations to assess convergence. For example, the authors could show differences between 1, 2, 5, 10, 20, and 60 minutes of temporal resolution to demonstrate how water levels respond to these changes. A similar approach can be taken for spatial resolution. During these comparisons, please avoid varying the bathymetry source simultaneously, as this would complicate the findings. When performing these analyses, provide an analysis that supports the findings. Why are water levels higher or lower with these settings?

Thank you for this good suggestion. While it would be highly valuable to explore the convergence of our findings across different temporal resolutions, the computational demands of such an analysis is very high and prevent us from performing this test. However, as the reviewer suggested, we have introduced an additional model configuration that isolates the effects of grid refinement without simultaneously altering the bathymetric dataset. These updates are now reflected in 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)."

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 (Xynthia case study)*

***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 also added a dedicated section in the results to analyze 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).*

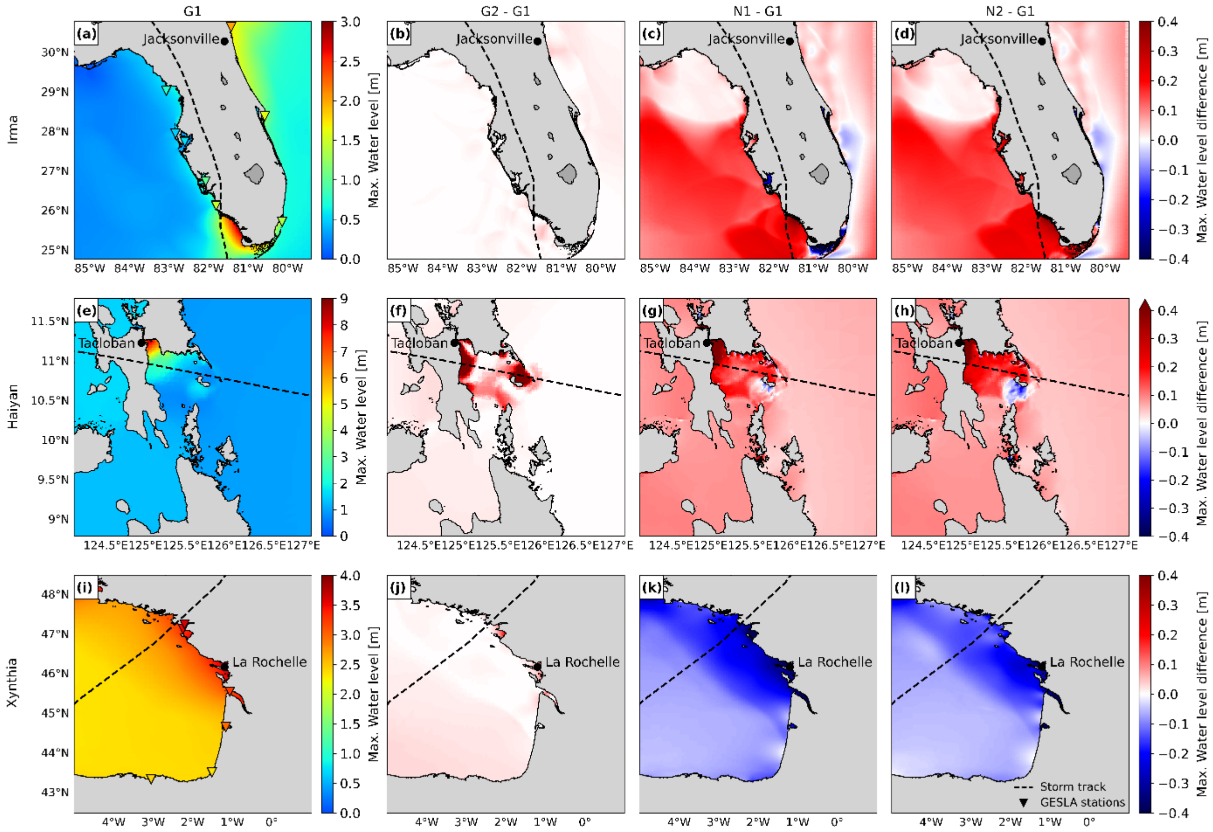

Figure 8. Maximum water levels for the three case studies, for G1 (panels a, e, i). Difference between the maximum water level for each specific model configuration (see Table 1) and G1. Panels a, e, i show observed maximum water level from tide gauge stations of GESLA. Difference in water levels for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

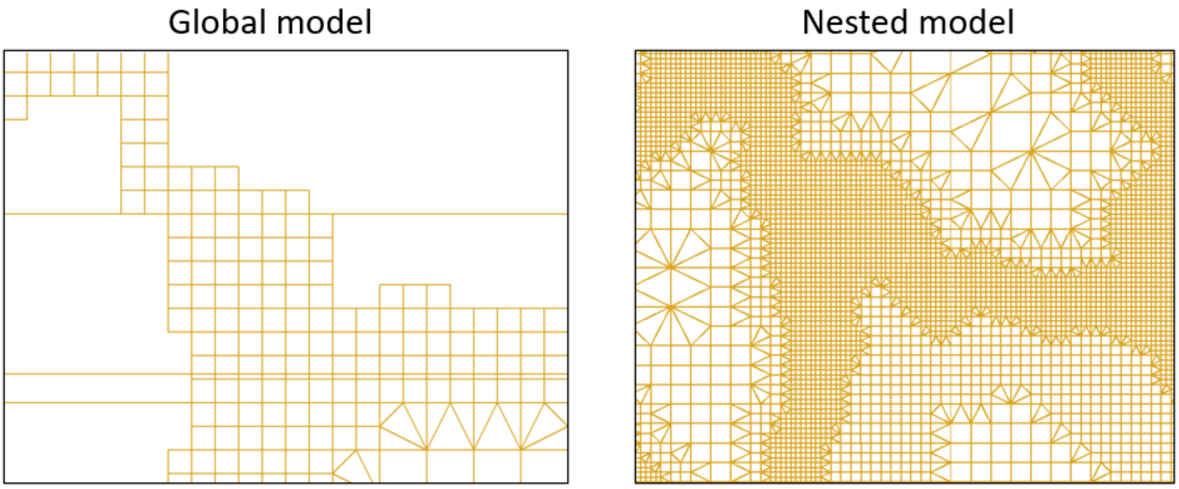

Figure A6. Close look at the unstructured grids of the global GTSM model with a grid resolution up to 2.5 km along the coast (left) and the nested grid of dynamic downscaling with a grid resolution up to 0.45 km along the coast (right), for case study Haiyan.

(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 10 panels e, j, o show that the model configuration N2 results in significant 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).

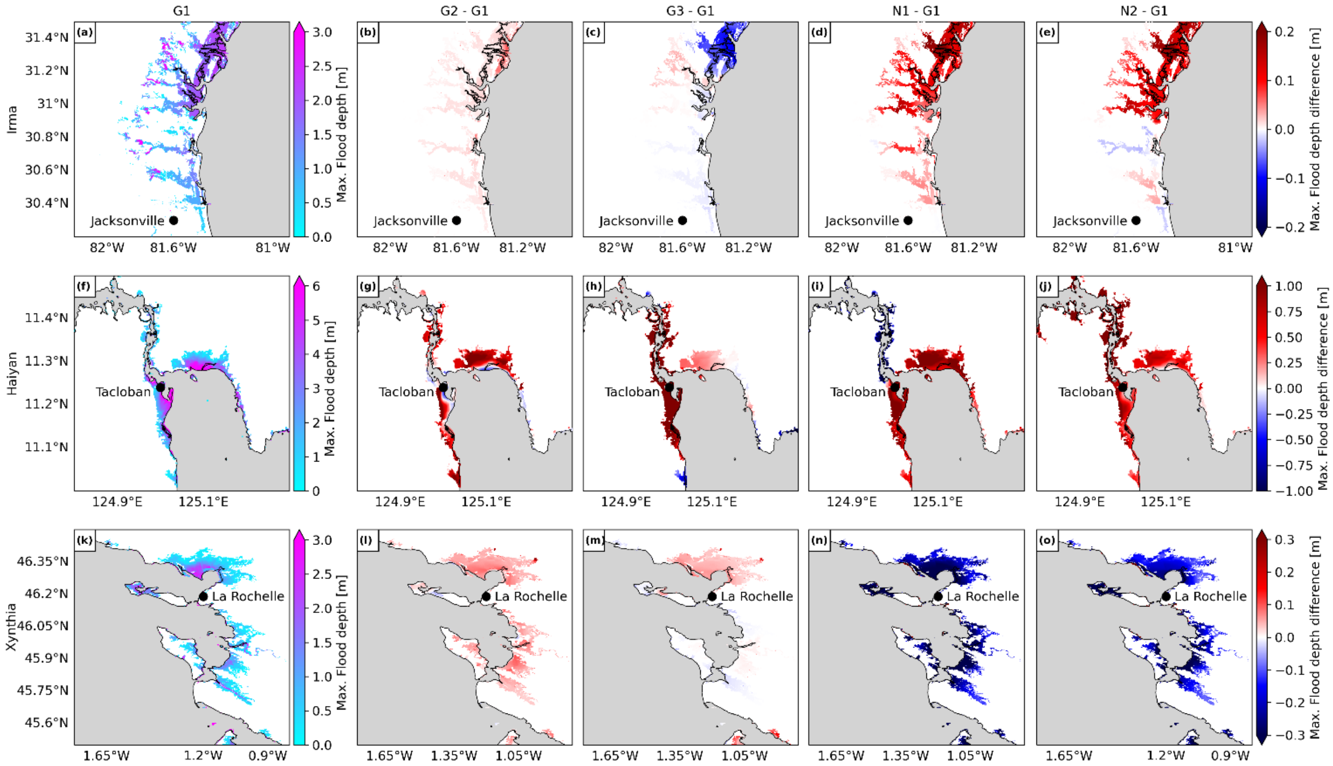

Figure 10. Panels a, f, k show the maximum flood depth for the default configuration G1, for each case study. Panels b, g, l show the difference between the maximum flood depth for the refined temporal output resolution configuration G2 and G1. Panels c, h, m show the difference between the maximum flood 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."

I am particularly skeptical about the dynamic downscaling/fully refined results. How do the authors explain a 40 cm increase in water level? It seems there might be a double-counting of the inverse barometer effect or another error. I do not believe that the entire Gulf of Mexico can have such a different water level based on minor model configuration changes. Could the authors provide more justification for these findings? To understand the results better, I recommend analyzing the time series first.

When validating the total water levels of Irma, Figure A2 presented above shows that specially for the southwest of Florida, while the default configuration G1 presents a slight underestimation of the peaks overall, the dynamic downscaling (N1 and N2) might overestimate those peaks slightly. Therefore, the ground truth is somewhere in between both modelling results, and it does not mean that G1 underestimated the peaks by 40 cm. When looking at table A1 presented above, both model configurations actually show similar results, with G1 having a RMSE of 0.28 m and a Pearson's correlation of 0.93, and N1 and N2 a RMSE of 0.27 and a Pearson's correlation of 0.93 and 0.92 respectively.

In this section, the authors use the word 'might' frequently. I suggest analyzing the results to test these hypotheses. For example, why are the results different for Haiyan with a 60-minute temporal resolution? One can demonstrate this by comparing water levels near the eye of the storm and further away, providing results rather than hypotheses.

In response to this and other comments from both reviewers, the result section was rewritten and significantly changed. We believe that these textual changes and additional analyses address the concerns of the reviewer. Specifically for Haiyan, we have included in the manuscript more results that help on the interpretation of the results:

> *(Lines 242 – 247): "For TC Haiyan (Fig. 8 panel f), the sensitivity of the water levels is significant. Water levels increase due to the temporal refinement up to 2 m along the coastlines where TC Haiyan made landfall, showing that 1-hourly resolution is too coarse to accurately capture the water level response. The cause for this is that TC Haiyan had a rapid intensification, and when modelling water levels at 1-hourly resolution we overlook the storm's peak, resulting in an underestimation of the maximum water levels. G2 however, can capture the peak of TC Haiyan more precisely (see Figs. A4 and A5)."*

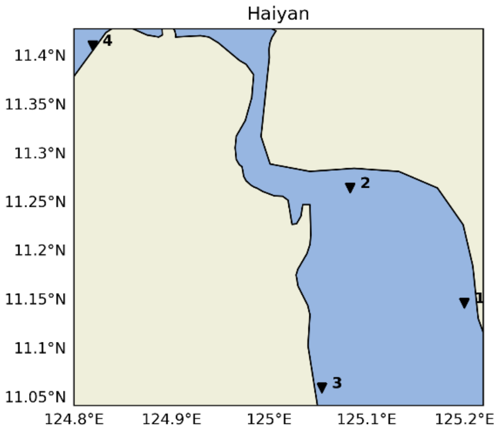

Figure A4. GTSM output locations for the case study Haiyan.

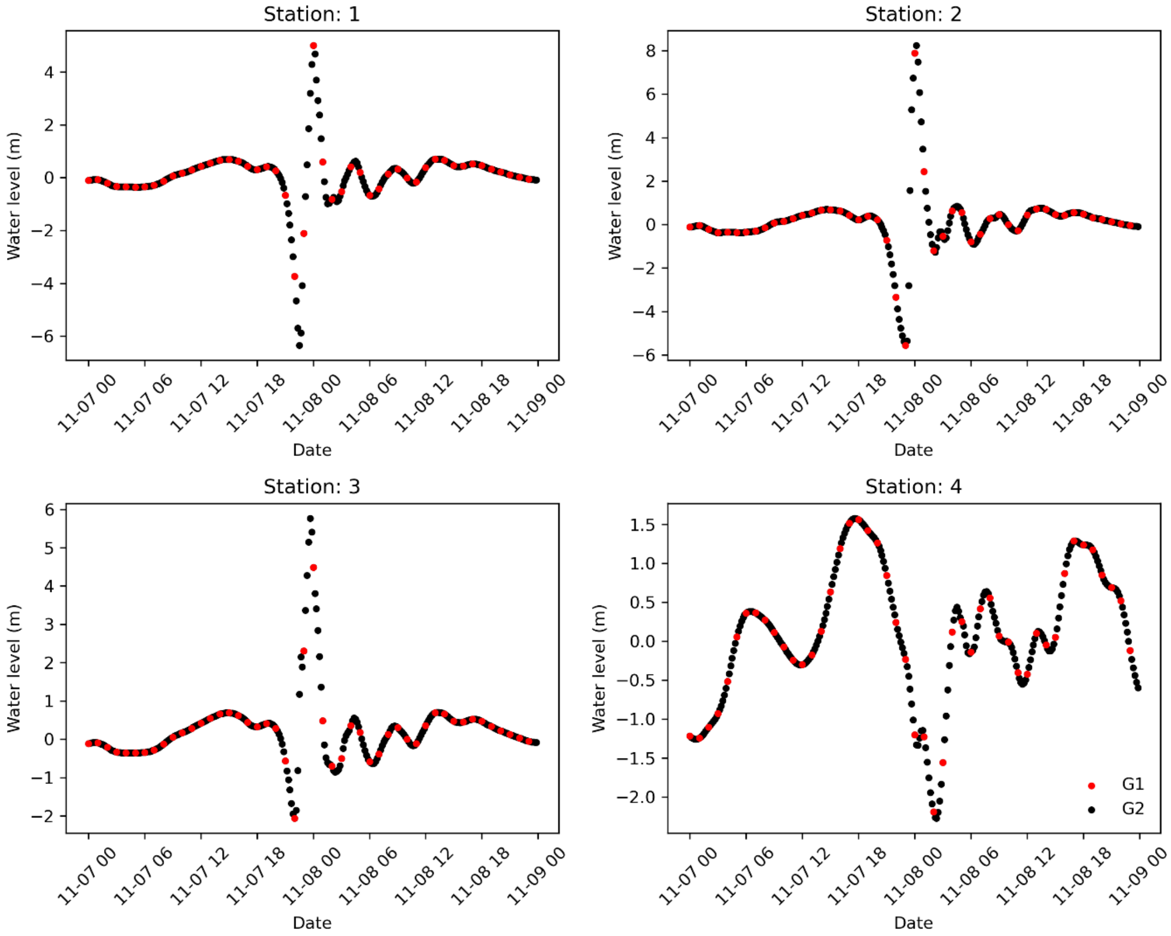

Figure A5. Haiyan total water level timeseries for the GTSM output locations provided in Fig. A3. Timeseries for the default configuration (G1) and the refined temporal output resolution configuration (G2).

In the flood section, the results are unconvincing. For example, during Irma, Jacksonville experienced severe flooding. In Figure 7 (a-d), the city appears unaffected. This is problematic. I suspect that topo-bathymetry is the cause, which brings us back to the challenges mentioned in the introduction that MOSAIC does not resolve. Demonstrating 1) accurate water levels near the city and 2) flood extents with more reliable US-based topo-bathymetry are essential to successfully model this case study.

Indeed the topography can play a key role in modelling flood depths and extents. Moreover, the flooding in Jacksonville was largely due to heavy precipitation. In this first version of our MOSAIC modelling framework we do not include multiple flood drivers and focus only on the surge as driver of coastal flooding. As a result of the validation suggested by the reviewer, we decided to update the DEM used for the case study Xynthia. In this region there were many dikes that prevented the water from travelling further inland. However, the DEM used before, FABDEM, could not resolve those. With the updated DEM from IGN, we can resolve better the dikes and obtain more accurate results. Nevertheless, the best approach for this would be to integrate, when possible, the flood protection measures in the hydrodynamic flood model. We have updated the manuscript as follows to integrate this:

(Lines 166 - 169): "For this study, we use GEBCO 2020 (15 arc seconds spatial resolution; (Weatherall et al., 2020)) as input dataset for the bathymetry and FABDEM (30 m spatial resolution; (Hawker et al., 2022)) as input dataset for the land elevation. Except for ETC Xynthia. For ETC Xynthia we use the 5 m resolution LiDAR-based DEM developed by the French National Geographic Institute (IGN) because it better represents dikes in the region, leading to better flood estimates than FABDEM (see Fig. A8)."

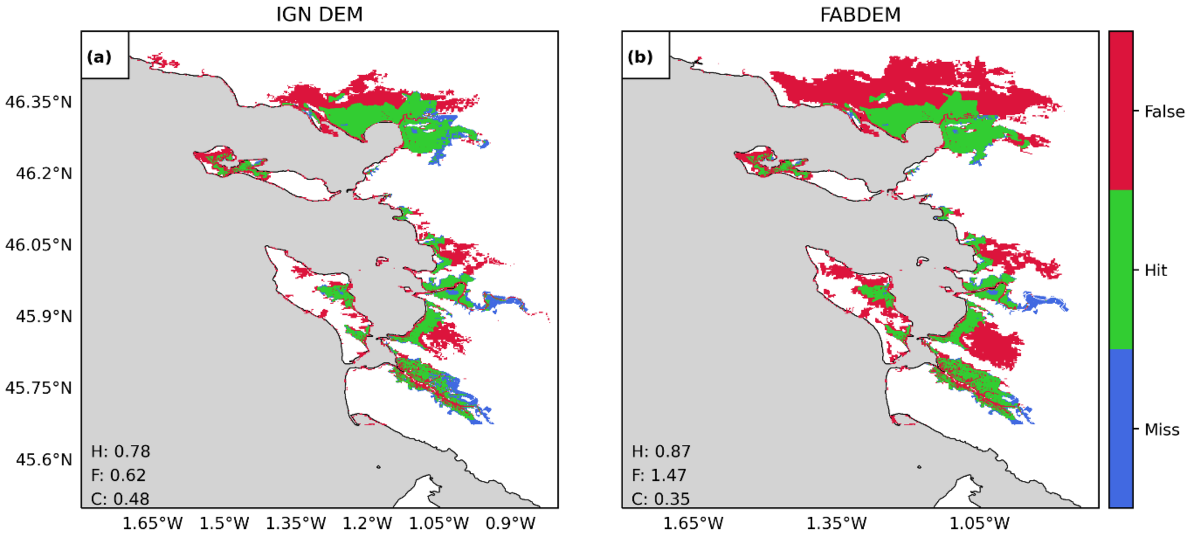

Figure A8. Validation of flood extents for the case study Xynthia against observed flood extents. The maps compare the modelled and observed maximum flood extents for a SFINCS model generated with ING's DEM (panel a) and FABDEM (panel b), where: green indicates flood areas correctly simulated; blue flood 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.

4. **Discussion**: I could not find the MOSAIC code on Zenodo, so I argue that this needs to be shared first before claiming it is 'automated and reproducible.' I also challenge the statement "enhance the simulation at the local scale by providing refined water levels." I have not seen evidence of this in the manuscript.

We have added the github link to MOSAIC in the manuscript. The datasets compiled during the study will be available on Zenodo upon acceptance of the paper:

*(Lines 488 – 490): "**Code availability***

The underlying code for this study is available on at https://github.com/Ireneben73/mosaic_framework (last access: 11 October 2024)."