Modeling surge dynamics improves coastal flood estimates in a global set of tropical cyclones

Check for updates

Thomas Vogt, Simon Treu ⓘ , Matthias Mengel ⓘ , Katja Frieler ⓘ & Christian Otto ⓘ ✉

Tropical cyclone-induced storm surge is a major coastal risk, which will be further amplified by rising sea levels under global warming. Here, we present a computational efficient, globally applicable modeling approach in which ocean surge and coastal inundation dynamics are modeled in a single step by the open-source solver GeoClaw. We compare our approach to two state-of-the-art, globally applicable approaches: (i) using a static inundation model to translate coastal water level time series from a full-scale physical ocean dynamics into inundated areas, and (ii) a fully static approach directly mapping wind fields to inundation areas. For a global set of 71 storms, we compare the modeled flooded areas to satellite-based floodplain observations. We find that, overall, the models have only moderate skill in reproducing the observed floodplains. GeoClaw performs better than the two other modeling approaches that lack a process-based representation of inundation dynamics. The computational efficiency of the presented approach opens up new perspectives for global assessments of coastal risks from tropical cyclones.

Tropical cyclones (TCs) are one of the most devastating categories of extreme weather events on earth, with their impact concentrated in coastal areas. TCs are connected to only 16% of the weather-related records in the international disasters database EM-DAT, but are responsible for 41% of the damages (1991–2020)[1]. They are so destructive because they encompass multiple hazards[2–7]: high-intensity winds, rain, storm surges, and associated compound floods. However, so far most global assessments on people exposed to TCs[8,9] as well as TC-induced economic damages[10–12] use wind as the only hazard predictor and neglect floods and rainfall.

To this end, we here focus on modeling the hazard of coastal floods, which were the primary cause of fatalities resulting from TCs over the 1963–2012 period in the USA[13]. Under climate change, the risk of TC-induced flooding is clearly expected to increase: sea level rise will amplify the risk of coastal flooding, and increasingly heavy precipitation will increase the risk of pluvial and fluvial flooding[14].

There are different approaches to estimate TC-induced inundated areas. First, full-scale ocean models such as ADCIRC[15,16], COMCOT[17], Delft3D FM[18,19], FVCOM[20,21], MIKE21[22,23], ROMS[24,25], or SLOSH[26,27] consider not only effects of the atmosphere, astronomical tides, and waves on hydrodynamics to model coastal water level time series; they also allow modeling inundation dynamics in the same modeling step. This has the advantage that they can account for the dynamic interactions of surges with astronomical tides, short surface gravity waves (wind sea and swell), or spatially varying surface friction characteristics over land[16].

Global flood risk assessments employing these models so far only used their coastal water level output and not their capability to model flood extents[28–31]. This contrasts the clear importance of dynamically resolving the surge dynamics[32–36].

Second, since flow solvers are often optimized to model either large-scale ocean dynamics, or small-scale estuaries and overland flooding, uncoupling the coastal surge from the coastal inundation dynamics is a common approach in coastal flood modeling[37]. Especially for multi-storm assessments, computational demand, and data requirements can be reduced when the full-scale physical (ocean) models are only employed to produce the time series of water levels along the coastlines, called hydrographs[38]. A computationally less expensive model can then use these as input for the actual inundation modeling. In global studies[39–41], it is common to use static (so-called planar, GIS-based, or bathtub-type) inundation models, as, for instance, the World Resources Institute's Aqueduct model[42]. These static bathtub-type approaches assume that any grid point is inundated if it has an elevation less than the extreme water level, usually attenuated according to its distance to the coastline.

Third, fully static and computationally inexpensive one-step approaches have been employed, especially for the assessments of global flood risks, as, for example, the storm surge module implemented in the open-source risk assessment toolbox CLIMADA[43]. They estimate the inundation height in each grid cell from wind speed, distance to coast, and topographical elevation using a linear relationship[44].

Potsdam Institute for Climate Impact Research, Potsdam, Germany. ✉e-mail: christian.otto@pik-potsdam.de

Finally, with machine-learning-based methods that do not rely on physical ocean models another computationally efficient alternative has emerged, recently. They can generate synthetic storm tide hydrographs[45,46], peak storm tides at a small selection of points[47–49], or even gridded flood extents[50–52] for risk analyses. However, the field of machine-learning-based flood area and depth assessment is in an early stage, and all the existing models are only calibrated for comparably small regions.

We here introduce a globally applicable, event-based, and dynamic one-step approach, building on the intermediate-complexity geophysical flow solver GeoClaw[53,54]. Solving the depth-averaged shallow-water equations (SWEs), GeoClaw allows us to seamlessly model TC-induced coastal water level time series and inundated areas along coastlines (Fig. 1). GeoClaw has an intermediate spatial resolution that is higher than global-scale ocean models but lower than most coastal inundation models. In contrast to static, not process-based (e.g., bathtub-type) approaches, GeoClaw accounts for the event-specific variation of wind and pressure over time as well as for the run-up of waves and overland flow around obstacles. GeoClaw's complexity is moderate compared to many flood models as it does not incorporate dynamic interactions with astronomical tides, short surface gravity waves (wind sea and swell), river discharge, or spatially varying surface friction characteristics over land. Astronomical tides cannot be imposed as dynamic boundary conditions in GeoClaw. This is why, for each simulation, we derive a constant offset from astronomical tides and add it to the zero water level in a preprocessing step.

We test the ability of GeoClaw in reproducing satellite-based observations of coastal inundated areas and compare its performance to two of the globally applicable model setups discussed above (Fig. 1): in the first model setup (GTSM + Aqueduct), we take a two-step approach. For that, we use coastal water levels calculated by the Global Tide and Surge Model (GTSM) version 3.0[55]—a depth-averaged global ocean dynamics model that dynamically simulates tides and storm surges using the unstructured Delft3D Flexible Mesh software[18]—to calculate inundated areas with the static bathtub-type inundation model of the World Resources Institute's Aqueduct

project[42]. In the second model setup (CLIMADA), we employ the fully static inundation model included in CLIMADA as a representant of light-weight approaches often used for large-scale risk assessments. Further, to test whether resolving the inundation dynamics improves the performance of GeoClaw, we combine GeoClaw with Aqueduct (GeoClaw + Aqueduct).

For model evaluation, we apply the models to a global set of 71 storms between 2000 and 2019 for which satellite-based observations of coastal inundated areas are available and compare their performance in reproducing these floodplains. We only include low-lying grid cells with a height of between 0 and 10 meters above geoid, and outside of permanent water bodies to reduce the share of flooded areas that are due to pluvial and fluvial floods (see "Methods" section for a detailed description of the observational datasets). In addition, we use coastal high water marks (HWMs), which are available for 11 US storms, for model evaluation (Fig. 2 and Supplementary Tables S1–S3).

We find that GeoClaw outperforms the other modeling approaches with regard to three common performance metrics. The inundated areas simulated by GeoClaw are generally larger but do not systematically overestimate the floodplains, while these are systematically underestimated by the other modeling approaches. To exclude the possibility that GeoClaw's better performance is merely a result of a more exact reproduction of coastal water level time series, we compare the simulated hydrographs as obtained by the different modeling frameworks with observational tide gauge data (available for 34 storms (Fig. 10a, b and Supplementary Table S5)). We find that these are better reproduced by the full-scale ocean model GTSM than by GeoClaw. These findings lead us to the conclusion that resolving the surge dynamics is the key driver of GeoClaw's better performance in reproducing coastal floodplains.

Results
Evaluation of modeled inundated areas
We first illustrate inundation for the different model setups through the examples of Hurricane Harvey (August 2017, Fig. 3) and Hurricane Rita
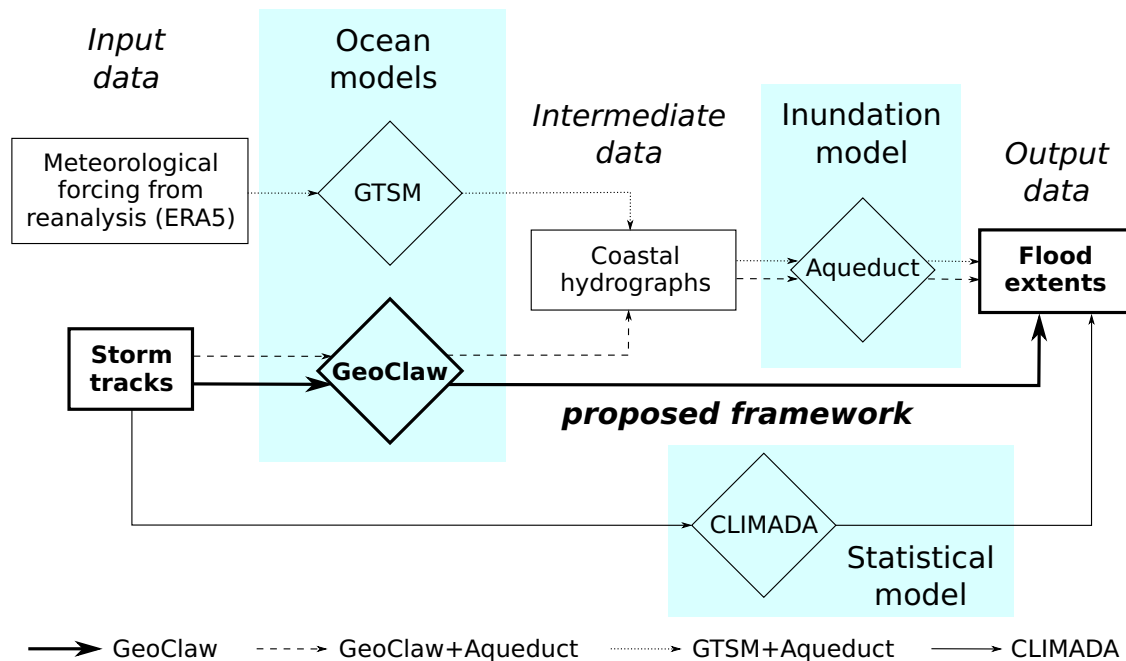

Fig. 1 | Comparison of the four considered modeling frameworks. Four modeling frameworks translating tropical cyclone storm surge into coastal flood extents are compared: (i) the shallow-water equation solver GeoClaw[54] employs storm track data to dynamically calculate coastal water level time series (hydrographs) and coastal flood extents in a single modeling step (GeoClaw, bold solid lines), (ii) the full-scale physical ocean model GTSM[55] is used to calculate coastal hydrographs from meteorological reanalysis data (ERA5[120]), which are then employed to calculate inundated areas with the static inundation model Aqueduct[42] (GTSM + Aqueduct, dotted lines), (iii) the fully static flood module implemented in CLIMADA[43] employs a statistical-surge relationship to translate storm track data directly into flood extents (CLIMADA, thin solid lines), (iv) to test the impact of dynamically resolving inundation processes, the static Aqueduct model is driven by GeoClaw's coastal hydrographs to calculate coastal flood extents (GeoClaw + Aqueduct, dashed lines). Arrows indicate the order in which the different models are executed.
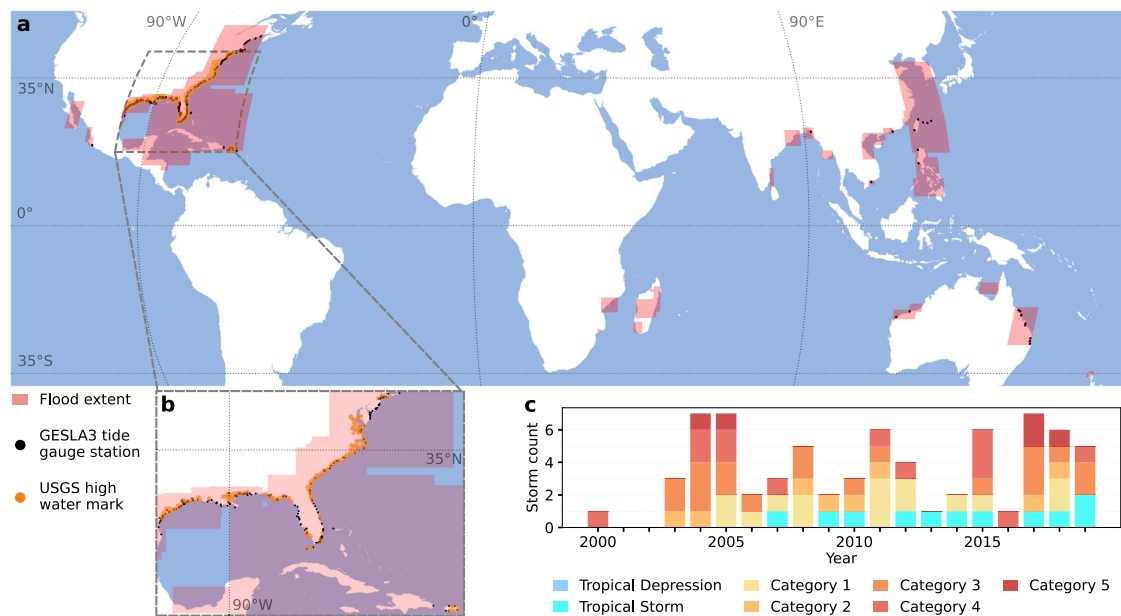
Fig. 2 | Overview of the observational validation data. a The validation data include 97 flood extents (red boxes) for 71 distinct tropical cyclone events. For 34 of the storms, tide gauge measurements are available. In total, 383 tide gauge time series at 213 distinct tide gauge stations (black dots) are used. **b** For 11 storms in the USA and Puerto Rico, a total of 1007 field measurements of inundation heights (high water marks; orange dots) are available. **c** The events cover the period 2000–2019, and all intensities of the Saffir–Simpson Hurricane Scale (color scale).

(September 2005, Fig. 4). Both storms made landfall on similar geographical regions in Texas and Louisiana and caused a significant storm surge. Extreme rainfall was the major cause of damages for Hurricane Harvey[56]. By contrast, the impact of Hurricane Rita was dominated by the devastating storm surge[57]. For both events, satellite-based flood extents are available from two different observational sources.

There are substantial differences between the inundated areas derived from satellite observations and all models. Also, the satellite observations differ between observational sources, though to a somewhat smaller extent. For Hurricane Harvey, two satellite-based products are available ("Methods" section): the RAdar-Produced Inundation Diary (RAPID)[58] and the Dartmouth Flood Observatory (DFO)[59], which estimate the flooded area to be 10,316 and 12,295 km^2, respectively. The flooded area calculated by GTSM + Aqueduct (494 km^2; Fig. 3f) is by one order of magnitude smaller than the observed flooded areas. Further, the size of the flooded area obtained with CLIMADA (1934 km^2) and GeoClaw (4866 km^2; Fig. 3c) is less than a fifth and less than half of the sizes of the observed areas. Also for Hurricane Rita, two satellite-based observations are available, one from the Global Flood Database (GFD)[60] (3673 km^2; Fig. 4a) and the other from the DFO (3239 km^2; Fig. 4b). The flooded areas obtained by GTSM + Aqueduct (1671 km^2; Fig. 4f) is less than half the size of the observed flooded areas. Further, the size of the flooded area calculated with CLIMADA (2659 km^2; Fig. 4d) matches the observations quite well, whereas the flooded area calculated with GeoClaw (7405 km^2; Fig. 4c) is nearly twice as big.

The three considered coastal inundation model setups (GTSM + Aqueduct, GeoClaw, and CLIMADA) match the observation more closely in the case of Rita than in the case of Harvey. Likely, one reason is that the flooding caused by Harvey was strongly driven by extreme rainfall not accounted for in the considered modeling setups[56]. (By contrast, Rita primarily caused devastating storm surges, whereas the amount of rainfall was comparably low[57]). However, the substantial overestimation by GeoClaw of the area flooded by storm Rita shows that missing fluvial flood components are not the only reason why mismatches between modeled and observed floodplains arise ("Discussion" section).

For Hurricane Harvey, the satellite-based products RAPID (Fig. 3a) and DFO (Fig. 3b) estimate flooded areas of 10,316 km^2 and 12,295 km^2,

respectively. The flooded area calculated by GTSM + Aqueduct is by one order of magnitude smaller (494 km^2; Fig. 3f) than the observed flooded areas. Further, the flooded areas obtained with CLIMADA and GeoClaw are less than a fifth (1934 km^2; Fig. 3d) and less than half (4866 km^2; Fig. 3c) as large as the observed areas, respectively.

Also for Hurricane Rita, two satellite-based observations are available, one from GFD (3673 km^2; Fig. 4a) and the other from DFO (3239 km^2; Fig. 4b). The flooded area obtained by GTSM + Aqueduct (1671 km^2; Fig. 4f) is less than half as large as the observed flooded areas. The size of the flooded area calculated with CLIMADA (2659 km^2; Fig. 4d) matches the observed ones quite well, whereas the flooded area calculated with GeoClaw (7405 km^2; Fig. 4c) is double the size of the observed flooded areas.

Foreseeably, the coastal inundation models match the observation more closely in the case of Rita than in the case of Harvey. This is because the flooding caused by Harvey was strongly driven by extreme rainfall not accounted for in the models[56]. By contrast Rita caused devastating storm surges, but the amount of rainfall was comparably low[57].

For both storms (Figs. 3c–f and 4c–f), the inundated areas calculated dynamically with GeoClaw reach further inland (Figs. 3c and 4c) than those obtained with both static, not process-based approaches (CLIMADA, Figs. 3d and 4d, and GTSM + Aqueduct, Figs. 3f and 4f). Especially, GTSM + Aqueduct inundates merely thin coastal strips. This suggests that the different characteristics of the flooded areas obtained with GeoClaw and the static modeling approaches may result from the dynamic representation of inundation processes in GeoClaw. (For instance, the static models neglect that the direction of the water flow is also driven by the wind and can change in the course of the storm, which may explain the higher penetration depth of the flooding obtained with GeoClaw.) To test this hypothesis, we calculate the coastal water levels with GeoClaw and the inundation with the static Aqueduct model (GeoClaw + Aqueduct). For both storms, this does not only reduce the size of the flooded areas compared to the full GeoClaw simulations (for Harvey and Rita the size of the flooded area is 619 km^2 (Fig. 3e) and 2324 km^2 (Fig. 4e) compared to 4866 km^2 (Fig. 3c) and 7405 km^2 (Fig. 4c), respectively, for the full GeoClaw setup). It also changes their characteristics, as, in this setup, only thin coastal strips are inundated, similarly to GTSM + Aqueduct. This is a strong indication that indeed the dynamic description of the inundation processes is a main cause for the
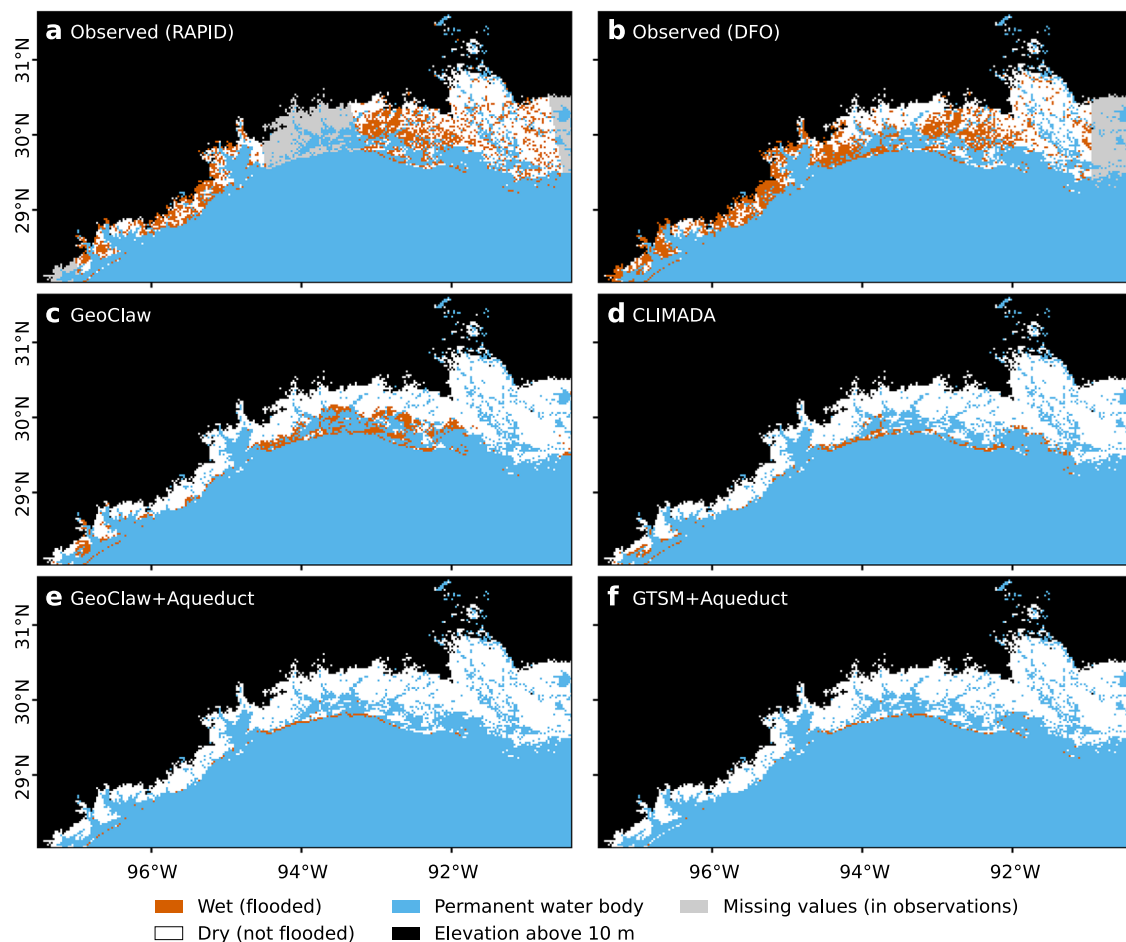
Fig. 3 | Observed and simulated flood extents for Hurricane Harvey (2017). The flood extents from two different satellite-based observations (**a** RAPID and **b** DFO) as well as from four inundation models (**c–f**) are shown, covering an 800 km stretch of coastline from Corpus Christi (Texas) in the west, via Galveston (Texas), to New Orleans (Louisiana) in the east. Some areas are marked as missing (gray) in the satellite-based products. Permanent water bodies (blue) as well as areas with an elevation of more than 10 meters above geoid (black) were excluded from the analysis. Wet (flooded) areas are marked in brown and dry (non-flooded) areas are marked in white. All products are reprojected to the same 1 km grid for comparison. In the model outputs, a grid cell is marked as flooded if the simulated flood depth exceeds 0.1 m.

different characteristics of the flooded areas calculated with GeoClaw compared to the static modeling approaches.

We next show that the inundated areas calculated with GeoClaw match the satellite observations systematically better than those from static inundation models. To this end, we first classify each cell of the flood maps for the full set of 71 TCs as either "wet" (positive) or "dry" (negative). We then use three performance metrics (scores, see "Methods" section) calculated from the areas classified as true positive, true negative, false positive, and false negative[61]: as the harmonic mean of precision and recall, the F1 score (with values between 0 and 1) is a metric for the accuracy of a prediction[62]. The F2 score (with values between 0 and 1) is the ratio of the area classified as wet by both (model and observation) divided by the area classified as wet by any of the two[63]. Despite being the predominant indicators of model performance, these scores are known to be biased in favor of overpredictions[64]. For this reason, we additionally consider the Matthews correlation coefficient (MCC), also known as (Yule) Phi coefficient. MCC is the Pearson correlation coefficient estimated for two binary variables (with values between −1 and 1). It is generally regarded as a more informative and true score if the class sizes (here the areas of "wet" and "dry" cells) vary[65] (see "Methods" section for a detailed discussion of the different performance indicators).

GeoClaw outperforms the static, not process-based approaches for all three performance scores for all 71 TCs (Fig. 5 and Table 1). This finding remains robust also if instead of the map-by-map score which results in a range of score values (colored boxes with black horizontal lines indicating median values in Fig. 5), we consider all floods at once which results in a single performance score for each model (total score; black stars in Fig. 5). Each flood map contributes equally to the map-by-map score, whereas the influence of each flood map on the total score increases with its geographical area.

Notably, the performance scores of all models considered in this study are comparably low and need to be interpreted in context. They cannot easily be compared to regional analyses or to studies about freshwater flooding. For example, all four approaches considered in this study have average F2 scores of <25%, while F2 scores of <30% are rather uncommon in the literature[66,67] (typical values range between 30% and 50%[68,69], and may even exceed 80% in some cases[70,71]). There are two main reasons: first, when modeling freshwater flooding over land, it is usually easier to define the affected area[67], while in the case of TC-induced coastal flooding it is more difficult to distinguish storm surge-induced flooding from pluvial and riverine flooding. Second, in contrast to many freshwater studies[68,72], we exclude permanent water bodies. (A detailed discussion and robustness checks can be found in the limitation section of the "Methods").

Inundated areas calculated dynamically with GeoClaw are larger than those obtained with static, not process-based models for the full set of the 71 TCs, confirming our findings for Harvey and Rita. To see this, we consider the ratio of true negative areas (TNR). A low TNR indicates that a prediction tends to overpredict, but underprediction is not penalized (classifying all cells as dry yields a perfect TNR of 100%). Further, we consider the bias score, which weighs overpredictions (too many wet cells) against
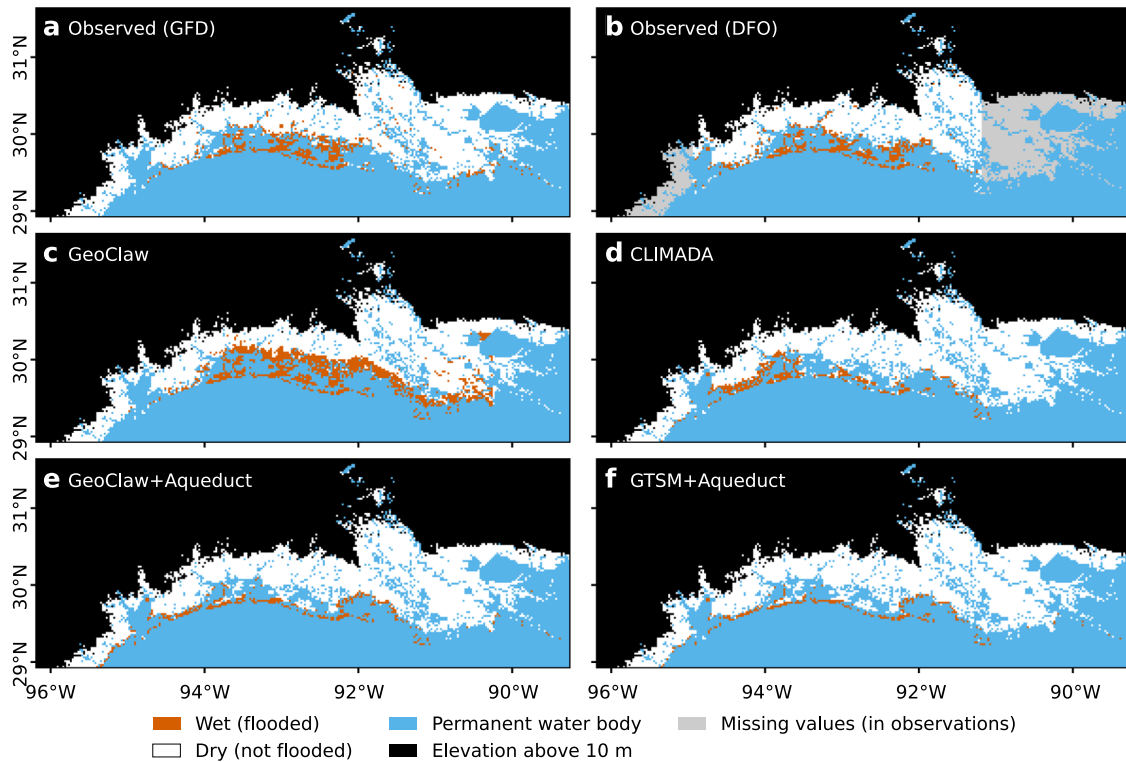
Fig. 4 | Observed and simulated flood extents for Hurricane Rita (2005). The flood extents from two different satellite-based observations (**a** GFD and **b** DFO) as well as from four inundation models (**c–f**) are shown, covering an 800 km stretch of coastline from Victoria (Texas) in the west, via Galveston (Texas), to New Orleans (Louisiana) and Gulfport (Louisiana) in the east. Some areas are marked as missing (gray) in the satellite-based products. Permanent water bodies (blue) as well as areas with an elevation of more than 10 meters above geoid (black) were excluded from the analysis. Wet (flooded) areas are marked in brown and dry (non-flooded) areas are marked in white. All products are reprojected to the same 1 km grid for comparison. In the model outputs, a grid cell is marked as flooded if the simulated flood depth exceeds 0.1 m.
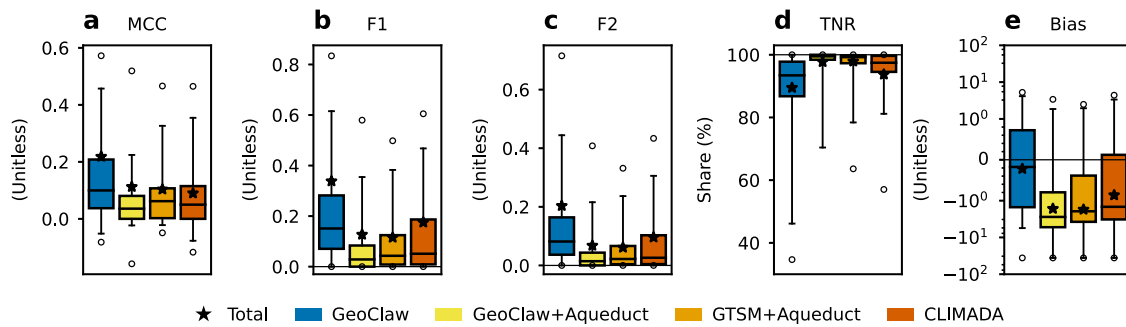

Fig. 5 | Comparison statistics for satellite-based flood extents. a–c Three metrics (higher is better) are used to express the overall performance of inundation models in reproducing the observed flood extents for 71 storm events. **d** The ratio of true negative areas (TNR) expresses the share of areas correctly classified as dry (higher is better)—an aspect that is not covered by the F1 and F2 performance scores (**b** and **c**). **e** The bias score expresses the tendency of a model to over- (positive values) or underpredict (negative values). Note the logarithmic scaling on the *y*-axis with linear scaling between −1 and 1. The evaluation can either be done for each flood map separately, resulting in a range of score values for each of the models (map-by-map score; colored boxes), or across all grid cells contained in all flood maps, resulting in a single performance score for each of the models (total score; black stars). The boxes denote the interquartile ranges with a horizontal black line for the median value, whiskers for the 95% intervals, and circles for the minimum and maximum values.

underpredictions (too many dry cells). A positive (negative) bias indicates a tendency to overpredict (underpredict). The TNR of GeoClaw (93%) is much lower than the TNR of the static inundation models (99% for GTSM + Aqueduct, and 97% for CLIMADA). The comparably high TNR of the latter suggests that the static inundation models underestimate the floodplains. This is also confirmed by the strongly negative biases of these models with median values ranging from −1.95 (GTSM + Aqueduct) to −1.47 (CLIMADA), whereas the bias for GeoClaw is much closer to zero (−0.18). Coupling the hydrographs obtained with GeoClaw with the static inundation model Aqueduct (GeoClaw + Aqueduct) yields a systematic

underprediction of inundated areas (99% TNR and −2.77 bias) which is similar to the other static approaches. As all models use the same topographic dataset, this indicates that the underestimation of inundated areas by the static models is caused by a systematic failure to represent some of the dynamic processes modeled by GeoClaw. We note that a tendency to underpredict is expected also for GeoClaw because it does not account for pluvial and fluvial floods while satellite-observed areas do not differentiate between different types of floods. Fully exploring the reasons for underprediction would require a model capturing all types of floods, which is beyond the scope of this analysis.

Table 1 | Performance indicators for the evaluation of flood extents (total scores and medians of map-by-map scores, separated by ★, with interquartile range in parentheses, Fig. 5)

Model	MCC	F1	F2	TNR (%)	Bias
GeoClaw	+0.22 ★ +0.10	0.34 ★ 0.15	0.20 ★ 0.08	89% ★ 93%	−0.22 ★ −0.18
	(+0.04 to +0.21)	(0.07 to 0.28)	(0.04 to 0.16)	(87% to 98%)	(−1.51 to +0.72)
GeoClaw + Aqueduct	+0.11 ★ +0.04	0.13 ★ 0.03	0.07 ★ 0.01	98% ★ 99%	−1.65 ★ −2.77
	(+0.00 to +0.08)	(0.00 to 0.08)	(0.00 to 0.04)	(98% to 100%)	(−5.27 to −0.80)
GTSM + Aqueduct	+0.10 ★ +0.06	0.12 ★ 0.04	0.06 ★ 0.02	98% ★ 99%	−1.74 ★ −1.95
	(+0.00 to +0.11)	(0.01 to 0.12)	(0.00 to 0.07)	(97% to 100%)	(−3.77 to −0.39)
CLIMADA	+0.09 ★ +0.05	0.18 ★ 0.05	0.10 ★ 0.03	94% ★ 97%	−0.86 ★ −1.47
	(+0.00 to +0.11)	(0.01 to 0.19)	(0.00 to 0.10)	(95% to 100%)	(−3.24 to +0.12)

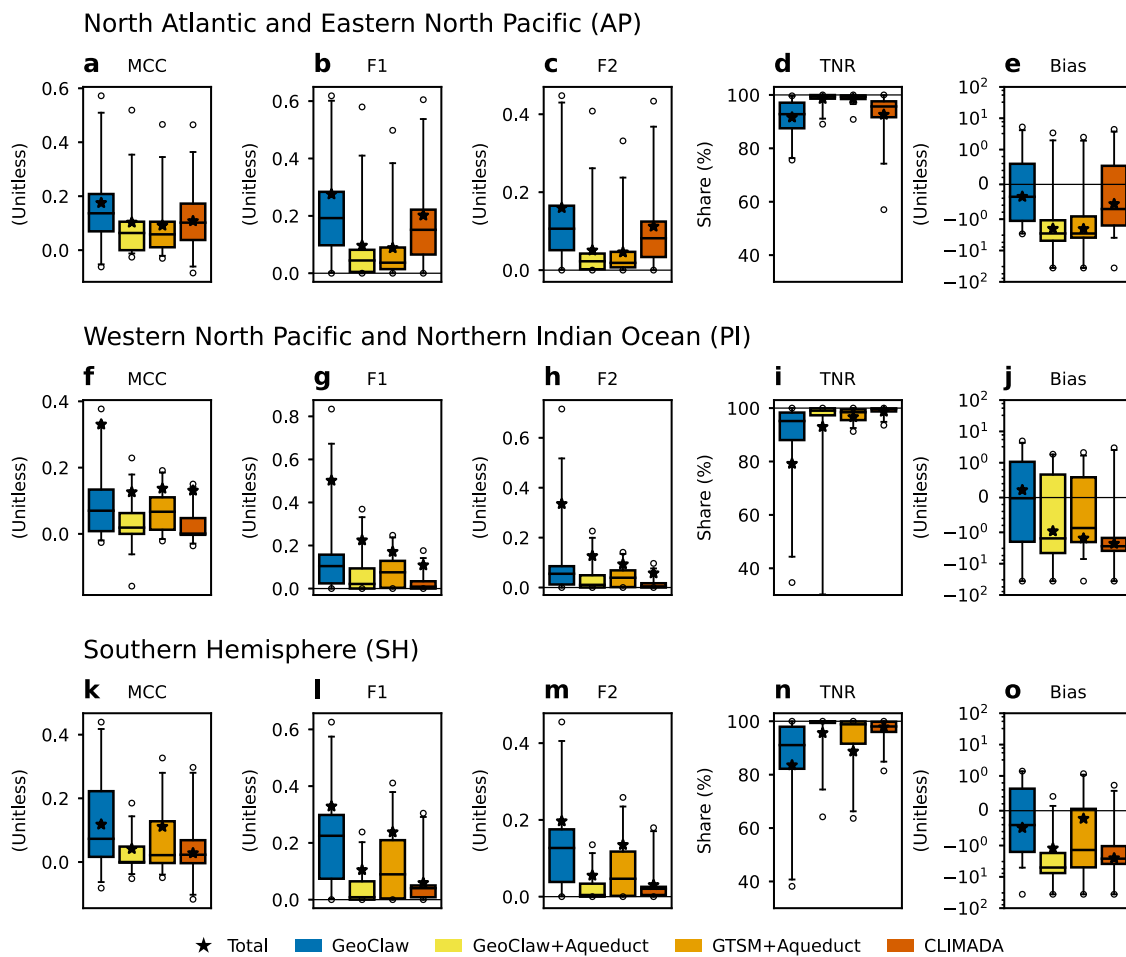For 95% intervals and minima/maxima, see Supplementary Table S4.

Fig. 6 | Evaluation of model performance by world region. The performance scores of modeled flood extents are shown for each of three world regions North Atlantic and eastern North Pacific (AP, **a–e**), western North Pacific and northern Indian Ocean (PI, **f–j**), and the Southern Hemisphere (SH, **k–o**). The colored boxes denote the interquartile ranges with a horizontal black line for the median value, whiskers for the 95% intervals, circles for the minimum and maximum values, and a black star for the total score. For a detailed description of the panels in each row, see Fig. 5.

We next investigate the performance per world region (Fig. 6). In our analysis, we distinguish three world regions with TC activity: the Southern Hemisphere (SH) as well as the North Atlantic and eastern North Pacific (AP) and the western North Pacific and northern Indian Ocean (PI). Without taking landfall and flood characteristics into account, approximately one-third of global TC activity is located in each of the three regions. However, in our flood extent data, SH is under-represented (20%, 14 storms), and AP is over-represented (46%, 33 storms), each by ~40%.

GeoClaw outperforms the static inundation models in all three regions: the scores of the three performance metrics, MCC, F1, and F2 are higher for GeoClaw. Further, the comparably higher TNR values of the static approaches suggest that these underestimate the observed flood extents more strongly than GeoClaw. This is also confirmed by the bias scores. For GeoClaw, the biases are negative for AP and SH and positive for PI. However, their magnitudes are much smaller than for the static modeling approaches. These have comparably strong negative biases in all three
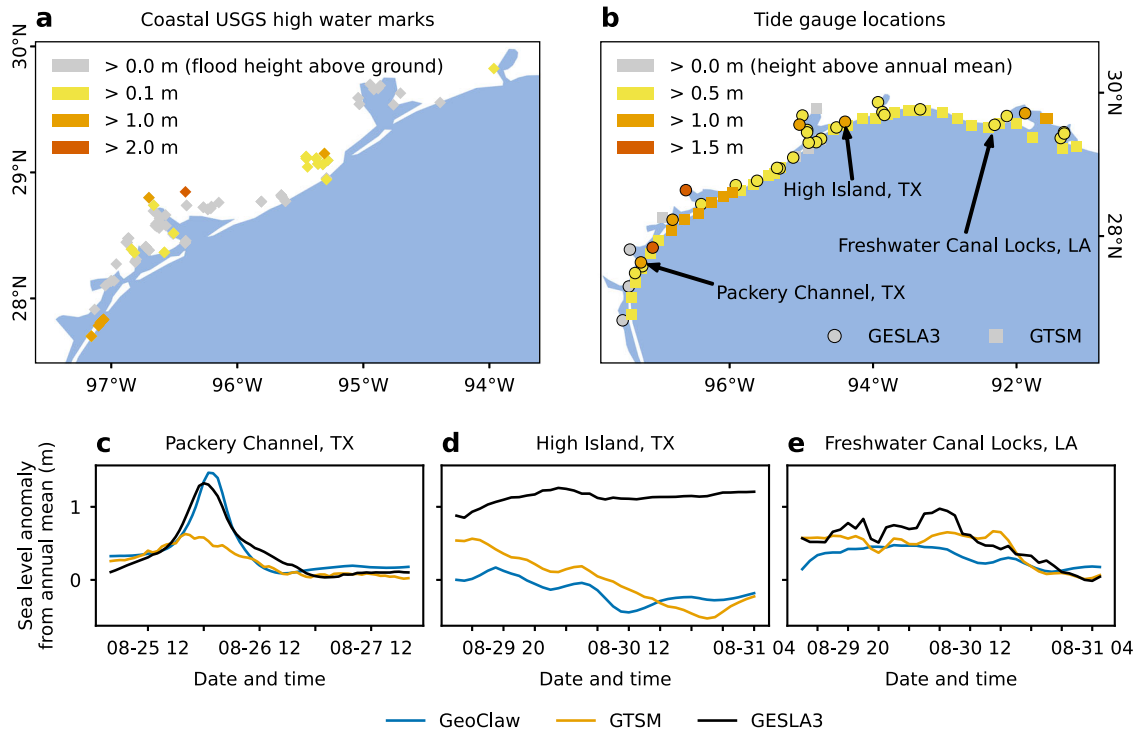
Fig. 7 | High water mark and tide gauge measurements for Hurricane Harvey (2017). a Inundation heights (color scale) above ground for 135 coastal high water mark locations (diamonds) along the Texas coast, measured in the aftermath of Hurricane Harvey in August 2017. **b** Maximum water levels (color scale) measured at 34 tide gauge locations (circles) and simulated at 42 additional output locations (squares). Measured and simulated water level time series (hydrographs) at Packery Channel (**c**), High Island (**d**), and Freshwater Canal Locks (**e**) tide gauge stations.

basins, indicating an underprediction of observed flood extents. A notable exception is GTSM + Aqueduct in SH, whose bias across all grid cells and maps is somewhat smaller in magnitude than the corresponding bias for GeoClaw. Generally, model performance is higher in AP than in the other two basins, except for GTSM + Aqueduct that performs in PI better than in AP.

Comparison of modeled inundated areas with high water marks
Satellites can both under- and over-prediction flood extents[67,73]. To gain a better understanding whether the satellite-based and modeled inundated areas over- or underestimate the actual storm surge-affected areas, we additionally compare them to field measurements of inundation heights (HWMs) from the US Geological Survey (USGS) Short-Term Network (STN)[74] (see "Methods" section). This dataset covers 1007 coastal HWMs for 11 TCs which made landfall in the USA out of the ensemble of 71 studied storms, including Hurricane Harvey and Hurricane Rita (Figs. 7a, 8a, and 9). Since HWMs come with a classification as coastal or riverine, we restrict our analysis to coastal locations to ensure that the recorded flood was predominantly caused by storm surge.

The tendency of satellite observations to underestimate flood extents is confirmed when considering HWMs. In fact, only 88% of the 1007 HWM locations are marked as wet in at least one of the satellite-based flood maps (white box in Fig. 9a). GeoClaw outperforms static approaches concerning the hit rate (HR) of flooded HWM locations. It estimates inundation for 92% of the HWM locations. CLIMADA (79%) and GTSM + Aqueduct (78%) score lower in predicting the flooded HWM locations (Fig. 9a and Supplementary Table S5). Like for the satellite-based flood extents, the agreement of the inundated areas calculated by GeoClaw + Aqueduct with the HWMs is lower (79%) than for the one-step GeoClaw approach (92%). This provides another indication that dynamically resolving the inundation processes is critical for a better representation of observed surge areas.

GeoClaw underestimates inundation heights, but with lower bias than the static models. In contrast to the satellite observations, the HWMs allow us to evaluate the inundation heights provided by the inundation models. All models differ from the HWM inundation heights by more than a meter (considering both under- and overestimation), but with quite different biases. For 896 among the 1007 HWMs, GeoClaw has non-zero inundation heights. GeoClaw somewhat underestimates the inundation heights (mean and median errors: 1.51 m and 1.16 m; interquartile range (IQR): 0.57–2.02 m; average bias: −0.49 m). The underestimation of inundation heights by the static inundation models (GeoClaw + Aqueduct: $N = 772$; mean and median errors: 1.80 m and 1.46 m; IQR: 0.69–2.42 m; average bias −0.88 m), GTSM + Aqueduct ($N = 762$; mean and median errors: 1.90 m and 1.42 m; IQR: 0.78–2.65 m; average bias −1.07 m), and CLIMADA ($N = 772$; mean and median errors: 1.61 m and 1.15 m; IQR: 0.49–2.15 m; average bias: −0.76 m) is even larger (Fig. 9b and Supplementary Table S5). Global Digital Elevation Model (DEM) biases are an important and often dominant source of uncertainty. Comparing the height information of the DEM used in this study with the ground elevation recorded at 666 of the 1007 HWMs shows that the elevations in the DEM tend to be too low (average bias of −1.44 m; IQR: −2.67 to −0.02 m; green box in Fig. 9b and Supplementary Table S5). For 50% of the HWM locations, the DEM deviates from the recorded elevations by more than 1.80 m. For 140 HWMs (21%), the overestimation of topographical heights in the DEM is higher than the actual inundation height. This means that the DEM error is so high that even with the true hydrograph, any static, bathtub-type inundation model would fail to predict flooding at these locations. This finding is only valid for those coastal areas in the USA, for which HWMs are available and cannot easily be translated to the global scope.

Evaluation of coastal water level time series
So far our analyses have provided several indications that resolving the inundation dynamics—as done by GeoClaw—could be critical to obtain larger floodplains which better agree with observed surge-affected areas than the floodplains obtained with static, not process-based models. However, there is still the possibility that GeoClaw produces larger
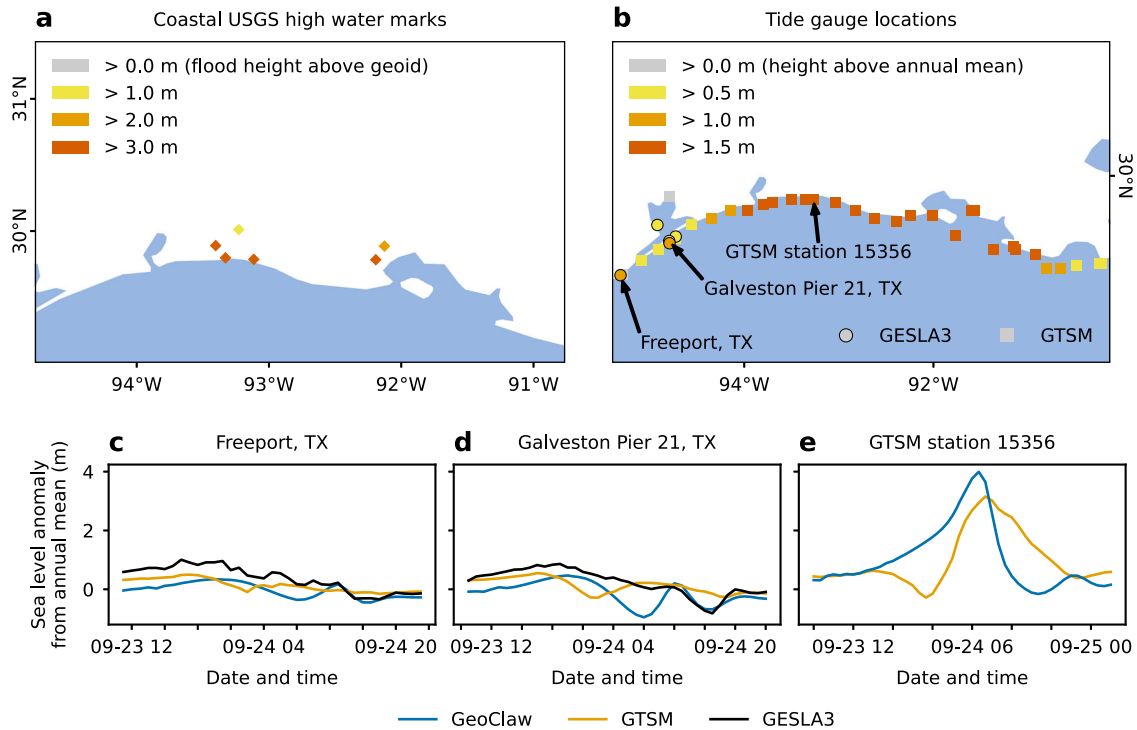
Fig. 8 | High water mark and tide gauge measurements for Hurricane Rita (2005). a Inundation heights (color scale) above ground for 6 coastal high water mark locations (diamonds) along the Texas coast, measured in the aftermath of Hurricane Rita in September 2005. **b** Maximum water levels (color scale) measured at 5 tide gauge locations (circles) and simulated at 32 additional output locations (squares). Measured and simulated water level time series (hydrographs) at Freeport (**c**), and Galveston Pier 21 (**d**) tide gauge stations, and at the GTSM output location with ID 15356 (**e**).
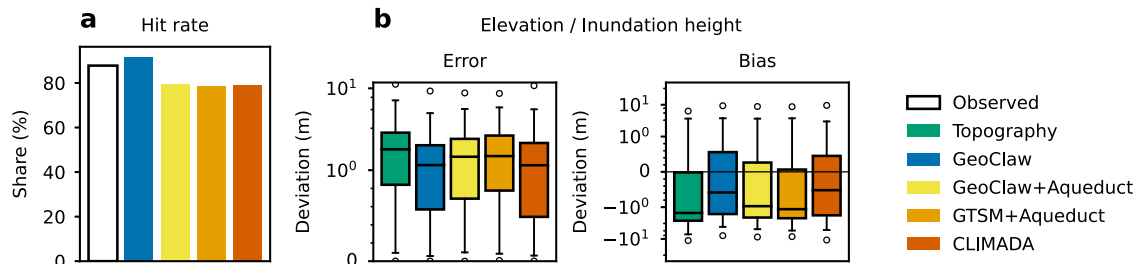

Fig. 9 | Comparison with high water marks (HWMs). a The share of HWMs for 11 storms in the USA and Puerto Rico where the satellite observations (white) and the inundation models (other colors) correctly classify a location as wet (hit rate). **b** Comparison of surface elevation and inundation heights recorded at HWMs to our digital elevation model (green box) and to the inundation models (other colors). Note the logarithmic scaling of the y-axis with linear scaling between −1 and 1. Colored boxes denote interquartile ranges with a black horizontal line for the median value, whiskers for 95% intervals, and circles for the minimum and maximum values.

floodplains because it overestimates coastal water level time series (hydrographs). We first test this hypothesis by evaluating the available hydrographs for Harvey and Rita, before assessing the hydrographs of all other storms for which data is available.

For Hurricane Harvey, GeoClaw does not systematically overestimate surge levels. We first compare observational and modeled hydrographs for hourly measurements of 34 tide gauge stations from GESLA version 3 (GESLA3)[75] (circles in Fig. 7b). The GeoClaw simulation, driven by a storm track from the IBTrACS archive (event number 2017228N14314)[76], underestimates the observed maximum surge height relative to the geoid by 0.33 m on average and the mean absolute deviation of maximum sea levels is 0.50 m. Similarly, GTSM underestimates maximum surge heights by 0.28 m on average, but the mean absolute deviations of maximum surge heights are significantly better (0.36 m). Which model performs better differs from station to station and sometimes both models are rather close and other times far off the observations (cf. tide gauge stations Packery Channel, Fig. 3c, High Island, Fig. 3d, and Freshwater Canal Locks, Fig. 3e). For

Hurricane Rita, the number of available tide gauge records (only five stations) is too low for aggregate statistics (Fig. 8b). Also, the few available tide gauge stations (Fig. 8c, d) are relatively far away from the areas that were most affected according to the GTSM simulations (Fig. 8e).

Observational water level time series from tide gauges are available only for 34 of the 71 TCs. In total, we can compare the simulated hydrographs to 383 observed hydrographs at 213 distinct tide gauge stations. GeoClaw does not reproduce coastal hydrographs with high accuracy (mean and median absolute deviations of maximum sea levels are 0.50 m and 0.42 m; IQR: 0.23–0.66 m; mean and median Pearson correlation coefficients of 0.50 and 0.66; IQR: 0.37–0.82 m; mean and median RMSE of 0.24 m and 0.20 m; IQR: 0.12–0.31 m; Fig. 10a, b and Supplementary Table S5). GTSM performs better in reproducing the observed hydrographs than GeoClaw (mean and median absolute deviations of maximum sea levels of 0.32 m and 0.26 m; IQR: 0.13–0.42 m; mean and median Pearson correlation coefficients of 0.65 and 0.82; IQR: 0.50–0.92; mean and median RMSE of 0.19 m and 0.15 m; IQR: 0.10–0.24 m). However, the improvement GTSM provides
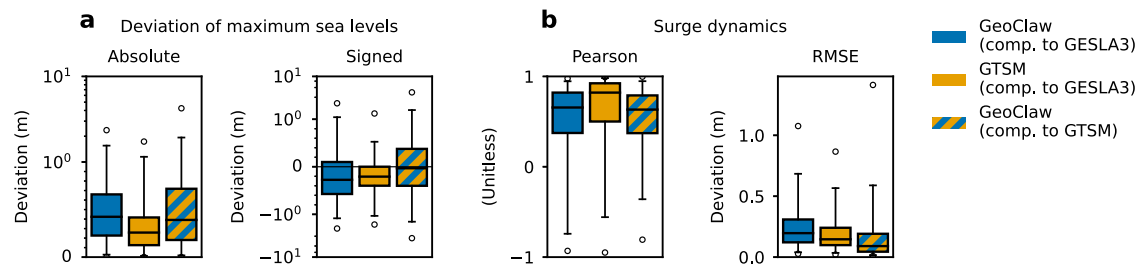
Fig. 10 | Comparison of observed with modeled hydrographs. We compare hydrographs observed at tide gauge stations with hydrographs simulated by GeoClaw (blue boxes) or by the state-of-the-art ocean model GTSM (orange boxes). We also compare the simulated hydrographs by GeoClaw with the GTSM simulations (boxes with blue and orange hatching). We compare (**a**) the absolute and signed deviations of maximum sea levels, and (**b**) the surge dynamics, using the performance metrics Pearson (higher is better) and RMSE (lower is better). The colored boxes denote interquartile ranges with a black horizontal line for the median value, whiskers for the 95% intervals, and circles for the minimum and maximum values. In (**a**), note the logarithmic *y*-axis scaling with linear scaling between −1 and 1.

compared to GeoClaw is moderate given that GeoClaw does not incorporate tidal dynamics or meteorological forcings other than the parametric TC wind and pressure fields. This might in part be due to the fact that GTSM is run at a lower resolution than our GeoClaw setup ("Methods" section). Moreover, GTSM is forced by wind fields from a reanalysis product that has been demonstrated to underestimate TC wind speeds[77,78]. The parametric wind fields used in our GeoClaw setup are often asserted to be a more reliable input for storm surge simulations[4,19,39,79]. Ideally, we would compare the models with identical resolution and wind forcing, but this is beyond the scope of this work.

Notably, both models underestimate surge heights; the signed differences of observed and modeled maximal surge heights have a bias of about 0.2 m (GeoClaw: −0.22 m; IQR: −0.57 m to +0.10 m; GTSM: −0.21 m; IQR: −0.40 m to +0.00 m). We thus infer that the larger surge-affected areas obtained with GeoClaw compared to the static approaches are a result of the dynamical modeling of the inundation dynamics and not due to a systematic overestimation of coastal hydrographs. Further, the comparably better agreement between the inundated areas calculated with GeoClaw and both satellite-based observations and HWMs shows that its ability to resolve the inundation dynamics overcompensates for its lower performance in reproducing observed hydrographs compared to the two-step approach of GTSM + Aqueduct.

Discussion

There is a clear need for better assessments of TC risks at different levels. They are needed to provide critical information and decision support for the Loss and Damage debate[80] at the international climate negotiations as well as by national actors responsible for the development and implementations of the National Adaptation Plans[81].

So far, global TC impact studies account for wind damages only. This may lead to an underestimation of TC risks for two reasons: first, already in the present climate, coastal inundation from storm surge are responsible for a large fraction of TC damage[13]. Second, while the evidence for a relation between climate change and the changes in TC wind speeds is rather weak and subject to large uncertainties, the risk of TC-induced coastal inundation is clearly expected to increase under climate change due to rising sea levels[82,83].

We here presented and evaluated an efficient intermediate-complexity framework based on the SWE solver GeoClaw[54] to dynamically model TC-induced coastal inundation. Having moderate data requirements and computational demands, it enables assessments on global scales. Considering a global set of 71 TCs, we showed that GeoClaw performs better than other available global modeling approaches in reproducing satellite-based floodplains: GeoClaw consistently outperforms a state-of-the-art global approach combining the computationally expensive ocean model GTSM with the static bathtub-type inundation model Aqueduct as well as the light-weight static approach implemented in CLIMADA[43].

To test whether GeoClaw's better performance in reproducing observed flood extents results from a better presentation of coastal water level time series (hydrographs) or its resolution of the inundation dynamics, we analyzed a subset of 34 TCs for which observational hydrographs are available. We found that GTSM systematically outperforms GeoClaw in reproducing the observed hydrographs. This leads us to the key insight that the dynamic inundation modeling by GeoClaw overcompensates its lower performance concerning coastal hydrographs. It also indicates that it is not necessarily required to go for ever higher mesh resolutions, more complex models, and high-resolution DEMs to improve floodplain modeling[84]. All these approaches demand a drastic increase in time, expertise, and computing power and may thus effectively postpone global assessments.

The presented approach is fully event-based and parallelized. This means that inundated areas can be calculated quickly, even for large numbers of TCs, depending on machine size. The main climate-dependent inputs are the TC wind fields. Large ensembles of synthetic storm tracks and associated wind fields can be efficiently generated for different future climates using TC emulators[9]. This allows for ensemble-based assessments of future inundation risks[10,12]. These can be more easily tailored to the needs of practitioners and policymakers than it is possible for approaches that rely on pre-computed global ocean model outputs[55]. On the other hand, our evaluation study also shows that both GTSM + Aqueduct and CLIMADA's flood module allow calculating meaningful inundated areas. This opens up the possibility to use the event-based open-source modeling framework presented in this paper for multi-model risk assessments.

Next to the assessment of future risk, the approach can be used for the attribution of the impacts of TC flooding to historical sea level rise[85]. For instance, our modeling framework was already used to estimate the contribution of climate change to the human displacements in the aftermath of Cyclone Idai, which made landfall in Mozambique in March 2019, accounting for sea level rise and changes in storm intensity due to global warming[5]. However, the main advantage of our approach is that it enables the global attribution of TC-inundated areas, since it allows calculating and comparing inundated areas with sea level rise (factual) and without sea level rise (counterfactual) for large global sets of storms.

The design of the evaluation experiments and the modeling framework are subject to several limitations (cf. "Methods" section). First, our approach currently does not allow for the simulation of compound flood hazards. Since extreme rain and coastal surge can lead to compound flooding, such an integration is necessary to assess the overall impacts of TCs. For instance, it was shown that compound effects played a crucial role in the case of Harvey[86,87] and are important to assess TC impacts in poorer regions that may quickly reach response limits[88]. The neglect of compound flood hazards may be also one main reason for the limited skill of the models to reproduce satellite-based floodplain observation. This indicates that further development of globally applicable compound flood modeling approaches is needed.

Second, GeoClaw is not able to integrate astronomical tides dynamically. Further, GeoClaw also only allows modeling one TC at a time, and cannot account for the influence of sequential TCs[89] or interactions with monsoon rainfall[90]. This is particularly important since sequential TCs are projected to occur more often under climate change[91], in line with an increase in multiple TC events[92]. Some of these issues affect the properties of the implementation and not of the conceptual design of GeoClaw, and could be added relatively easily to GeoClaw in the future.

Third, also the satellite-based observations come with some important shortcomings, limiting their suitability for storm surge model evaluations. Most importantly, the inundated areas may not only result from storm surge but also from pluvial floods and river floods. This renders it a priori difficult to decide whether the—on average—larger inundated areas obtained by GeoClaw are indeed more realistic than those of the static approaches or whether GeoClaw "accidentally" performs better because it overestimates the extents of the storm surges and thus partially replaces the missing flood drivers. To limit this risk, we restrict our analysis to low-lying (<10 m) areas and consider performance metrics that account for both over- and under-predictions. Note, however, that this reduces, but does not remove the influence of confounding drivers completely since the low-lying areas still experience pluvial and riverine flooding as in the case of Hurricane Harvey[87] (Fig. 3).

Fourth, satellite-based flood maps tend to underestimate the flooded areas because satellites can only detect water bodies that persist until the satellite flies over the given area[93–95]. Moreover, even in cases where timely coverage is available, there is evidence that satellite-based flood extents can suffer from both under- and over-predictions[67,73]. Due to the potential underestimation of the inundated areas by satellite-based products, the inundation models could underestimate the low-lying inundated areas even more strongly than already suggested by our performance analysis (Fig. 5 and Table 1). Finally, comparing the flood extents as obtained from different satellite observations, we find that the disagreement between observational products is quite substantial (Supplementary Fig. S2 and Supplementary Note 1). This indicates that satellite-based observations do not reach the quality of ground-based observations, which is why we complement them by coastal HWMs wherever possible.

The presented coastal inundation modeling framework based on GeoClaw represents an important building block and step forward toward more reliable global risk assessments. By adding coastal inundation risks to wind-related risks, it allows for more comprehensive comparisons of TC risk profiles across regions, globally. In the "hotspots" regions identified by the global risk analysis, more complex physical compound flood models that fully account for surge dynamics as well as coastal, riverine, and fluvial flooding can be used for refined risk assessments[96–98]. In the future, these may be complemented by machine-learning-based models, which have recently shown promising results for local flood risk assessments[50–52].

Methods
Overview
The four approaches to estimate coastal flood extents from TCs that are compared in this study are a combination of the four models GeoClaw, GTSM, Aqueduct, and CLIMADA (Fig. 1). We describe the models and their respective input and output data in the following subsections. Here, we provide an overview of the approaches and data.

Our proposed model setup with GeoClaw produces flood extents from TC storm track data. The CLIMADA modeling approach uses the same input data and produces the same kind of output data, but does not resolve any flow dynamics. The other two approaches GTSM + Aqueduct and GeoClaw + Aqueduct are a combination of a dynamic ocean model (GeoClaw or GTSM) with the static inundation model Aqueduct: coastal hydrographs are computed as an intermediate data product that is used by Aqueduct to translate peak storm tides at the coast into flood extents using a bathtub-type method.

Model runs of GeoClaw, Aqueduct, and CLIMADA have been conducted for this study. For GTSM, we used publicly available model outputs[55].

GeoClaw, Aqueduct, and CLIMADA use the same topo-bathymetric data set, while GTSM uses GEBCO[99]. Moreover, while GeoClaw and CLIMADA use wind and pressure forcing derived from TC track data using parametric models, the meteorological forcing for GTSM is taken from reanalysis data.

Both dynamic models (GeoClaw and GTSM) are based on the two-dimensional depth-averaged SWEs. Both use finite volume methods to solve the equations, but while GeoClaw uses structured (rectangular) grids, unstructured (curvilinear) grids are used in GTSM. In both models, the spatial resolution increases in coastal areas, starting from a 25 km resolution over the open ocean. GTSM uses a fixed, pre-computed grid, with a resolution of up to 2.5 km, while GeoClaw uses adaptive mesh refinement (AMR) to increase the grid resolution locally during run time to up to 0.25 km. GTSM uses the globe as a single model domain and runs for the whole period 1979–2017. For GeoClaw, independent runs encompassing 48 h each and a storm-specific geographical domain are set up for each individual storm. While GTSM models only ocean cells with shearless free-slip boundary conditions at coastlines, GeoClaw seamlessly models coastal inundation processes for land cells. In GTSM, astronomical tides are included through tide-generating forces. GeoClaw is not able to model tidal forcing dynamical. Instead, we apply a static correction of the zero water level, based on pre-computed astronomical tides. In both models, wind and pressure are incorporated as forcing terms in the momentum equations, but different wind and pressure data sets are used as input data.

While GeoClaw uses a physics-based description of bottom friction to model the coastal inundation processes, the static models Aqueduct and CLIMADA use a simplified concept of friction that assumes a linear decrease in flood height with distance from the coast. The rate of decrease (resistance factor) is constant for CLIMADA, and depends on the topographic terrain features for Aqueduct. The computational demands for the dynamic models GeoClaw and GTSM are so high that a high-performance computing cluster is necessary to run the models at the 1 km resolution used in this study. Even then, computing the dynamics for a single storm event can take several hours. The static models CLIMADA and Aqueduct, on the other hand, can be run on a laptop computer with run times of less than a minute for CLIMADA, or several minutes for Aqueduct.

Compared to regionally refined two-step approaches with dynamic ocean and inundation models, the data requirements and computational demands of our GeoClaw-based approach are moderate. However, it still requires substantially more time and expertise than CLIMADA. While Aqueduct is also comparably simple, it takes hydrographs from dynamic ocean models as an input. The most important features that make our proposed GeoClaw-based approach manageable: the software is open-source and available under a free license; the input data are freely available for all global regions; the setup and configuration of GeoClaw can be done in Python, even though the core of GeoClaw is written in Fortran; numerical and model parameters do not require case-specific adjustments or calibrations; the mesh is automatically adapted to the study region and event, and does not need to be pre-computed and adjusted manually; when applying the approach to large ensembles of storms, only the most intense ensemble members require long computational times, ensemble members with low intensity require less computational time; a GPU-accelerated version of the code is available, further decreasing run times if GPUs are available.

Dynamic surge and inundation from GeoClaw
We model surge heights and inundated areas using the open-source geo-physical flow solver GeoClaw that solves the depth-averaged SWEs[53]. We configure GeoClaw's AMR feature to start with a 0.25° (25 km) grid over the open ocean and refine the mesh in coastal areas to up to 9 arc-seconds (180–270 meters, depending on latitude), with five intermediate refinement levels. The resolution over the open ocean thus agrees with the one used in the GTSM setup (see below). In coastal areas, the resolution is higher than common global ocean model configurations: for recent global hindcast data sets, GTSM was configured with 2.5 km resolution[55,100], and SCHISM[28] with 2 km resolution at the coastlines. As our model setup aims at an output

resolution of 1 km (30 arc-seconds), we selected the lower bound for the input resolution to be slightly lower (9 arc-seconds) for numerical stability. Note that GeoClaw locally adjusts the resolution automatically for each run, and 9 arc-seconds is only a lower bound.

The topo-bathymetric data used with GeoClaw is a combination of three global DEMs. We use SRTM15+ (version 2.3) globally as the source of bathymetry with a resolution of 0.5 km[101], and the global CoastalDEM (version 2.1)[102] dataset for coastal areas. CoastalDEM has a focus on coastal areas and comes at a resolution of 90 m. For the land area that is not covered by CoastalDEM, we use MERIT-DEM[103]. Since all three DEM products (SRTM15+, CoastalDEM, MERIT-DEM) are already provided relative to the same vertical datum (the Earth Gravitation Model from 1996, EGM96), we overlay them without vertical adjustment. We first resample SRTM15+ from 15 to 3 arc-seconds resolution to fit the resolution of MERIT-DEM and CoastalDEM. Then, we fill missing values in MERIT-DEM with SRTM15+ (this mostly affects the bathymetry) and overlay the result with CoastalDEM outside of permanent water bodies (as provided by ESRI on https://hub.arcgis.com/content/e750071279bf450cbd510454a80f2e63/). Finally, we resample the combined result to a resolution of 30 arc-seconds (~900 m). This smoothes out transitions between the different data sets while being sufficiently accurate in the context of our inundation models that are configured to generate results at this resolution. While high-resolution LiDAR DEMs are available for most of the US coastal areas and several other regions (e.g., Australia), we decided to use global DEMs everywhere for consistency. The applicability with globally harmonized data sets is a primary concern in the design of our approach, since we aim at global assessments of climate risk. Also, the handling of LiDAR DEMs can be challenging since they usually cover comparably small regions and DEMs for neighboring regions are often not harmonized. We used the combined land-ocean data set SRTM15+ for bathymetry instead of the pure ocean data set GEBCO[99] because the seamless transition from ocean to land is important in our one-step approach. CoastalDEM is used because it is the best available global DEM for coastal modeling[104]. MERIT-DEM is used to fill small gaps in CoastalDEM since CoastalDEM does not cover the whole land area, and SRTM15+ has lower resolution than MERIT-DEM. A comparison with other choices of global DEMs would be an interesting topic for future research[105].

In a preprocessing step, we divide each storm into temporal periods (modeling periods) of at most 48 h. Due to the adjustment of zero water levels according to local astronomical tidal conditions (see below), the period may not be too long since tidal conditions change over time. On the other hand, the length must be sufficient for the spin-up of the wind-induced flow dynamics. Those parts where the distance of the storm to the closest coast is larger than twice the radius of maximum winds are excluded. GeoClaw is then started for each of the modeling periods separately. For each modeling period, the computational domain of the simulation is chosen to be large enough to accommodate the storm track together with a buffer of 2.5 times the radius of the outermost closed isobar.

Along the boundaries of the computational domain, GeoClaw uses extrapolation (non-reflecting or outgoing) boundary conditions that let the model waves from inside the model domain pass through the boundary without reflection. There is currently no way to impose dynamic water level boundary conditions according to astronomical tides. Instead, we set the zero water level (for the water body at rest) to the maximum astronomical tide attained in the center of the affected coastal area during each modeling period. Coastal areas are taken to be affected if the distance to the storm eye is not more than twice the radius of maximum winds. The astronomical tides are taken from the FES2014[106] simulations, referenced to a geoidal vertical datum using the gridded satellite altimetry product by AVISO[107]. Since the satellite altimetry is relative to the geoid model GOCO05s, we further converted the heights to EGM96, the geoid model used in the global DEM datasets (see above). We use the maximum since our comparison with tide gauge measurements showed that the simulated surges tend to underestimate tide gauge observations even with this zero water level configuration. To demonstrate the sensitivity of our model setup to this setting,

we show results for three different assumptions on the zero water level in Supplementary Figs. S7–S9. The flow dynamics are forced by wind speed W and air pressure P_A. GeoClaw derives this information on-the-fly from observed parameters along a storm track data set using the Holland 1980 parametric wind and pressure model[108]:

$$W = \sqrt{\left(\frac{r_{\max}}{r}\right)^B \cdot W_{\max}^2 \cdot e^{1-(r_{\max}/r)^B} + \frac{(r \cdot f)^2}{4}} - \frac{r \cdot f}{2}, \qquad (1)$$

$$P_A = P_c + (P_n - P_c) \cdot e^{-(r_{\max}/r)^B}, \qquad (2)$$

$$B = \frac{\rho_{\text{air}} \cdot W_{\max}^2 \cdot e}{P_n - P_c}, \qquad (3)$$

where P_c, $W_{\max}$, and $r_{\max}$ are central pressure, maximum wind speed, and the radius of maximum winds, as included in the storm track, f is the Coriolis parameter, B is Holland's fitting parameter[108], $P_n = 1013$ hPa is ambient pressure, and e is Euler's number. As storm track input we use observational data from IBTrACS, the most comprehensive global dataset of historical TC activity[76]. IBTrACS collects information reported by the WMO Regional Specialised Meteorological Centers (RSMCs) and by agencies in Shanghai and Hong Kong. For each of the 71 TC events in our study, we extract the IBTrACS data about the following variables, following the methodology in ISIMIP3a[109]: time, location of the storm center, central pressure, maximum 1-min sustained wind speed, environmental pressure, radius of maximum wind speeds, and radius of the outermost closed isobar. Apart from the parametric wind and pressure fields, no meteorological forcing is applied in our setup.

The interaction of the water flow with the air above and the surface below are implemented as friction terms in the momentum equations of the SWEs: for wind friction, the Garratt wind drag law[110], an approximation of the Charnock equation[111], is used:

$$C_w = \min(2 \cdot 10^{-3}, (0.75 + 0.067 \cdot W) \cdot 10^{-3}). \qquad (4)$$

The bottom friction is implemented as a Manning term:

$$C_f = \frac{g \cdot n^2}{h^{4/3}}, \qquad (5)$$

where g is the gravitational constant, and n is the "Manning coefficient", which is set to 0.050 on land, and 0.025 off shore.

As output, maximum inundation heights in coastal areas are stored on a 30 arc-second grid (in meters above ground, and above geoid). For this purpose, the internal height values, which are a primary variable in the SWEs, are interpolated from the internal mesh of varying resolution to the regular 30 arc-second grid. In addition to the inundation maps, we configure GeoClaw to store time series of flood heights at predefined output locations, according to the GESLA3 tide gauges and CoDEC-ERA5 output locations (see below). The surge dynamics are stored at the (varying) temporal resolution of the GeoClaw simulation run, but are resampled by taking hourly averages for the evaluation. This agrees with the resolution of GESLA3 tide gauge measurements (see below).

For our analysis, we ran GeoClaw on nodes of a high-performance cluster with 64 GB of RAM and 16 cores at 3.4 GHz each (i.e., 16 OpenMP threads). The run times (wall times) for a single TC event ranged from 30 s to 21 h with an average run time of 2 h and 34 min. Half of the TC events were completed in under 1 h, and only 23% of the TC events needed more than 3 h. Note that the run times differ a lot between TC events because GeoClaw uses AMR so that the computational demand depends strongly on the intensity of an event. Many events in our analysis are Category 1 or weaker (Fig. 2). For comparison, a two-step modeling framework based on LISFLOOD-FP at 90 m resolution has been reported to require 24 h when running on 200 cores[112]. Note that there is a GPU implementation of

GeoClaw[113] that was not used in this study, but would further speed up the process for operational applications (3.6–6.4 times faster than a 16-core CPU).

Choice of parametric wind field model

The use of the parametric wind field model has a considerable impact on the surge dynamics[114,115]. Currently, GeoClaw uses the Holland 1980 radial wind field model in its standard implementation[108]. This model can produce unrealistically large storm systems, which can be traced back to the derivation of the wind field from the pressure field using a gradient balance relationship[116]. In the updated 2010 version of the Holland wind field model[117], the representation of the radial wind profile is derived from an exponential and power-law composite function. The updated version can be better empirically constrained if observations of the wind field outside the radius of maximum wind are available. This leads to a better description of the wind field outside the hurricane eye wall in the calibrated Holland 2010 model[118].

In a case study for Hurricane Michel, Yann et al. assessed the impact of the choice of the wind field model on the surge dynamics by driving a hydrodynamic model with the 1980 and the 2010 version of the Holland model[116] with the 2010 model fitted to observed outer wind radii. The authors showed that the choice of the wind field model impacts the surge dynamics but the performance of both model setups in reproducing maximum water levels were similar.

Since for most of the considered storms, no observations of wind radii outside the storms' eye walls are available, we here decided to keep the Holland 1980 model. However, it would be of interest for future studies to analyze how the choice of the wind field model impacts on GeoClaw's performance.

GTSM dynamic ocean model

We use pre-computed outputs of the GTSM from the extreme sea level dataset CoDEC-ERA5[55] that is freely available from https://zenodo.org/records/8322750[85]. The data come as hourly-resolved sea level time series at 18,719 output locations, equidistantly located every 50 km along the smoothed global coastlines. We refer to the literature for a detailed description of the model setup[55]. Here, we provide a summary for the convenience of the reader.

The GTSM[55] uses the Delft3D Flexible Mesh software to solve the depth-averaged SWEs on a global unstructured mesh with spatially varying grid resolution which varies between 25 km in the deep ocean and 2.5 km in coastal areas. Only ocean and no land grid cells are part of the model grid so that coastal inundation processes are not modeled. At the coastal boundary, the movement of particles tangential to the coastline is assumed to be shearless (free-slip boundary condition). Astronomical tidal forcing is applied in the form of tide-generating forces in the momentum equation of the SWEs[119]. The model is forced with wind and pressure fields from the European Reanalysis (ERA5)[120].

CLIMADA static inundation model

We apply the static inundation model included in the open-source risk analysis toolbox CLIMADA[43] to the historical events considered in this study. A detailed description of the model can be found in the official documentation (https://climada-petals.readthedocs.io/en/latest/tutorial/climada_hazard_TCSurgeBathtub.html). Separately in each grid cell, this model estimates inundation heights using a linear relationship with three physical predictors: wind speed, distance to coast, and topographical elevation. While the model lacks statistical and physical justification and has not been validated with observational data so far, it is still used for probabilistic impact and risk assessments due to its computational simplicity as well as low requirements on data availability[44].

Compared to other static inundation models, the CLIMADA model does not require maximum surge heights as an input but it implicitly estimates surge heights from maximum wind speeds. The wind speeds are derived from IBTrACS[76] storm track data sets using a parametric model that

is implemented as part of CLIMADA[108,121]. For storm track and topographical elevation, we use the same data used in our GeoClaw setup (see above). The only other input data set is the gridded Distance to the Nearest Coast dataset that is freely distributed by NASA. We configure CLIMADA to compute outputs on the same rectangular grid with 1 km resolution that is used in our GeoClaw setup.

Aqueduct static inundation model

We apply the static, bathtub-type inundation model that is published as part of the World Resources Institute's Aqueduct project[42], and available as open-source software from https://github.com/Deltares/aqueduct-coastal-flooding/tree/py38 under the terms of the GNU General Public License version 3. This model derives inundation heights in inland coastal areas from maximum surge heights given at points along the coastline. Even though the model is static, it implements a concept of friction which is called "resistance factor"[122], meaning that flood heights decrease with the distance to coast, depending on surface properties. The resistance factor is $0.5~\text{m km}^{-1}$ on open terrain, with higher values for higher topographical elevation. For grid cells that are frequently inundated by permanent water bodies, the resistance factor is reduced proportional to water occurrence statistics, so that the resistance factor vanishes in grid cells that are permanently part of water bodies. For water occurrence statistics, we use the Copernicus Global Surface Water raster data for 1984–2019 at 30 m (0.9 arc-seconds) resolution[123]. For each grid cell, the surface water occurrence, i.e. the frequency with which water was present in the study period as a percentage between 0 and 100, is provided.

Previous studies used the Aqueduct model to derive inundated areas from coastal peak storm tide outputs of the GTSM[39,42,122,124]. Its performance has not been validated with observational data so far. We apply the Aqueduct model to peak storm tides at the CoDEC-ERA5 output locations computed with GTSM and GeoClaw (see above).

Performance indicators

For the evaluation of the simulation results, we compare the areas marked as flooded by either one of the satellite-based or modeled flood extents using several performance metrics. Since the prediction of flood extents is a binary classification problem (with classes "wet" for positive and "dry" for negative), the scores are expressed in terms of areas classified as true positive (n_{11}), true negative (n_{00}), false positive (n_{01}), and false negative (n_{10})[61].

The MCC[125,126], also known as (Yule) Phi coefficient, is the Pearson correlation coefficient estimated for two binary variables:

$$\text{MCC} = \frac{n_{11} \cdot n_{00} - n_{01} \cdot n_{10}}{\sqrt{(n_{11} + n_{01}) \cdot (n_{11} + n_{10}) \cdot (n_{00} + n_{01}) \cdot (n_{00} + n_{10})}}. \quad (6)$$

The F1 score[62] is the harmonic mean of *precision* (positive predictive value, PPV) and *recall* (HR):

$$\text{F1} = \frac{2 \cdot \text{HR} \cdot \text{PPV}}{\text{HR} + \text{PPV}}, \quad (7)$$

$$\text{HR} = \frac{n_{11}}{n_{11} + n_{10}}, \quad (8)$$

$$\text{PPV} = \frac{n_{11}}{n_{11} + n_{01}}. \quad (9)$$

Note that we set the F1 score to the value 0 if $n_{11} = 0$. The F2 score[63], also known as critical success index (CSI), threat score (TS), flood area index (FAI), or Jaccard index, is the ratio of the area classified as wet by both (model and observation) divided by the area classified as wet by any

of the two:

$$F2 = \frac{n_{11}}{n_{10} + n_{11} + n_{01}}. \qquad (10)$$

The TNR score[127] (ratio of TNR), also known as specificity, is the ratio of areas marked as dry by both divided by the areas marked as dry in the observation:

$$TNR = \frac{n_{00}}{n_{00} + n_{01}}. \qquad (11)$$

In this study, the bias score[127] is defined as the log-fraction between wet areas, predictions compared to observations:

$$Bias = \log\left(\frac{n_{11} + n_{01}}{n_{11} + n_{10}}\right). \qquad (12)$$

Note that the Bias score can take the value $-\infty$ in cases where the model does not predict any flooding. However, this does not affect the aggregation of map-by-map scores since we do not take averages but only quantiles which are robust for this kind of outlier.

MCC, F1, and F2 quantify the overall quality of the prediction with a single number that takes into account true and false positive and negative classifications. While the MCC score has a minimum value of -1, F1 and F2 range from 0 to 1. For all three scores, a higher value is better, with a value of 1 denoting an exact fit. The TNR score takes a value of 100% in case of a perfect fit, but contrary to the previous scores, a higher TNR does not imply that the fit quality is better, in general. A low TNR indicates that a prediction tends to overpredict, but underprediction is not penalized: classifying all cells as dry yields a perfect TNR of 100%.

Data on satellite-based flood extents

The main source of validation data for our study is satellite-based flood extents. We consider data from three flood extent databases:

1. In the GFD, flood extents have been derived from MODIS satellite measurements[60]. Each of the 913 flood extents is linked to an entry in the flood list of the DFO[59], covering the period 2000–2018. Using the description, geolocation and date of the entries, we matched TC events from IBTrACS to 61 of the available flood extents. We removed 6 events from the selection since we found no overlap in coastal areas between the storm geometry according to IBTrACS and the flood extents provided by GFD.
2. There is a continuously updated collection of rapid response flood maps published directly by DFO[59], based on MODIS satellite measurements. The collection is not provided in machine-readable form, but only for human access through a web interface, as RGB color images, most of them not properly georeferenced. Among those, we identified 61 maps that are related to 44 TC events in IBTrACS, covering the period 2001–2019. We manually converted these maps to a machine-readable georeferenced format for further analysis. After that, 15 maps were removed from the selection due to a missing overlap with IBTrACS storm geometries. In previous studies, DFO flood extents were used for a selection of inland flood events in Nigeria (2012)[128], Mozambique (2007)[72], and for flood events along the Brahmaputra River (2012)[129]. Other than that, the DFO flood extents have not been used for similar flood extent validation studies.
3. Finally, the near real-time system for inundation maps named RAPID is based on synthetic aperture radar (SAR) devices on-board earth-orbiting platforms[58]. A selection of maps produced with RAPID is publicly available in machine-readable form. Among those, we identified 9 flood maps that were related to 6 TC events in IBTrACS, covering the years 2016–2019, all located in the North Atlantic region. One of the maps was removed from the selection as covering only non-coastal areas. Due to the SAR-based measurements, the flood maps only cover parts of the affected areas.

Altogether, we selected 97 flood maps that are related to 71 TC events in IBTrACS, covering the years 2000–2019. For the comparison of the simulated flood extents with observational data (Fig. 5), we aggregated the satellite-based flood extents from their original resolution (between 1 and 27 arc-seconds, 20–850 m) to the resolution of all simulation outputs (30 arc-seconds, 900 m). During aggregation, a grid cell is classified as flooded if at least one of the underlying grid cells in the original resolution was flooded. We only include grid cells with a height of between 0 and 10 meters above geoid, and outside of permanent water bodies. A threshold of 10 m is common to define the low-elevation coastal zone in impact assessments[40]. To define permanent water bodies, we use the Copernicus Global Surface Water raster data for 1984–2019 at 30 m (0.9 arc-seconds) resolution[123]. For each grid cell, the surface water occurrence, i.e. the frequency with which water was present in the study period as a percentage between 0 and 100, is provided. For our analysis, we aggregate the raster data to the resolution of the model outputs (30 arc-seconds) using averaging and include only grid cells for which the occurrence frequency exceeds 5%.

The size of the flood maps and the size of the wet and dry areas within each map varies a lot from map to map. The size of the coastal area (including both wet and dry) covered by the 97 satellite-based flood maps differs by orders of magnitude from map to map, ranging from 170 to more than 75,000 km^2 (Fig. 6a). In total, 200,000 km^2 are classified as wet. For each map, this is between 0 and more than 27,000 km^2, with a mean and median wet area of 2060 and 480 km^2 (66% interval: 100–3400 km^2). Seventy-five extents have smaller than average wet areas. Together, 8 extents account for more than 50% of the areas observed as wet. For 17 among the 97 flood maps, the wet area is <0.05% (100 km^2) of the total wet area. Note that restricting our evaluation to extents with a medium-size observed area of between 100 and 1000 km^2 does not change the results of our analysis significantly (Supplementary Fig. S1a–e). For our analysis, we process the simulation outputs of the four approaches, a grid cell is marked as flooded if the flood depth exceeds 10 cm. This threshold is appropriate for the low-resolution outputs that we consider here. In studies with high-resolution LiDAR elevation data, a lower threshold is also common, e.g. 1 cm[69]. Note that omitting the threshold does not change the results of our analysis significantly (Supplementary Fig. S1f–j).

Data from tide gauge measurements

We use the hourly tide gauge data provided by GESLA3[75], a consistent global data set of tide gauge measurements. There are duplicate tide gauge stations in GESLA3, e.g. most stations operated by the National Oceanic and Atmospheric Administration (NOAA) are also listed in the database by the University of Hawaii Sea Level Center (UHSLC). Since GESLA3 collects data from both providers, several stations are listed twice in GESLA3. Therefore, we selected for each flood map and tide gauge location the tide gauge provider in GESLA3 with the lowest number of missing values. Furthermore, we restrict to the stations within the geographical area of each flood extent that were operational during the time of landfall of the storms. Since the flood extents are often larger than the actual extent of the TC, we restricted the analysis to the tide gauge locations that lie within the IBTrACS storm geometries. We further excluded tide gauge stations with large reporting gaps in the period of the TC event. More precisely, at most 1 hourly data point may be missing in each 24 h period.

Using the satellite altimetry (see above), we shifted the water levels in the tide gauge records so that they are relative to the geoid. For that, we extract the annual means from the gridded satellite altimetry product at the location of each tide gauge station. Note, however, that this correction is subject to major uncertainty, since the satellite altimetry product has a resolution of only 0.25° (~25 km). Furthermore, the satellite altimetry can only provide height information for grid cells that are off shore. Many tide gauge stations lie very close to the coast or even in narrow estuaries, so that their location is often in grid cells that are not off shore according to the satellite altimetry. In those

cases, we consider the average value of neighboring grid cells. We excluded stations from the analysis where neither the containing nor any of the directly neighboring grid cells has altimetry information.

After that, there remained 383 records from 213 distinct tide gauge stations, covering 34 of the 71 storm events. Only the time span where a TC signal is to be expected in the surge was used in the evaluation. For that, we estimated an hourly wind speed time series at each tide location using the wind field model included in CLIMADA (see above). The time span for the analysis was then chosen to start 12 h before the first time the wind speed exceeded tropical storm strength (17.5 m/s), and end 12 h after the last time the wind speed was above that threshold.

Data from field measurements (high water marks)

For a set of 11 of the 71 events in our analysis, HWMs from the USGS Flood Event Viewer[74] are available. This product is only available for the USA. It consists of field measurements by volunteers and trained USGS hydrographers, usually directly following a flood event[130]. The observers take note of mud, seed, debris, cut or wash lines on the inside and outside of buildings (in residential areas) and on open terrain features such as trees, shrubs, grasses, bluffs or river banks. Most HWMs are documented with a photography, a numerical value indicating the height above ground in feet, and the vertical uncertainty in one of six categories of uncertainty ranging from Excellent (±0.05 ft) to Very Poor (more than ±0.40 ft).

In previous studies, the HWMs have been used for the validation of a rapid storm surge forecasting framework[131], and for the validation and comparison of event-based flood inundation mapping services[132]. In general, field measurements such as the HWMs used here are commonly taken as a complement to conventional water-level records in assessing model performance[133].

HWMs are only available for the conterminous United States and only for selected TCs. Since the horizontal and vertical datums vary among HWMs, we excluded HWMs with missing datum information. We reprojected the remaining HWMs to the EGM96 vertical datum, the geoid model used in the global DEM datasets (see above). For 11 of the 71 storm events for which flood extents are available, we identify 2171 HWMs that lie on land within the IBTrACS storm geometries and at a topographic height of at most 10 meters (according to our DEM data set). For our main HWM analysis (Fig. 9), we only use the 1007 HWM locations that are classified as coastal (for the same analysis with only riverine locations, see Supplementary Fig. S4).

Limitations of the performance indicators

Since we have GTSM hydrographs on a fixed set of output locations, we could not compare the simulated GTSM hydrographs with observed hydrographs at the exact locations of the GESLA3 tide gauge stations. Following previous studies[55,85], we compared each GESLA3 hydrograph to the simulated GTSM hydrograph at the closest available output location. However, we found several examples where the tide gauge stations and output locations were separated by estuaries or small islands. In those cases, the difference in the hydrographs might be dominated by the difference in location. Therefore, we excluded those tide gauge stations from the analysis, where the distance to the next GTSM output location is larger than 10 km. The specific choice of 10 km was a trade-off between the number of tide gauge stations we remove and the error due to the distance between GTSM grid cell and tide gauge station. In the literature, The mean difference between modeled GTSM and observed water levels is reported to be 0.19 m for the 1 in 10 year peaks[55]. This agrees well with the analysis of TC hydrographs in this study: we find that the mean difference between modeled GTSM and observed GESLA3 peak water levels is 0.26 m across the 71 TC events. To complement the aggregate statistics in Fig. 10a, we also illustrate the comparison of maximum water levels using a scatter plot in Supplementary Fig. S5.

In the comparison of HWMs, we evaluate the absolute inundation heights (above geoid) instead of the inundation heights above ground because the inundation heights above ground are missing for 45% of the

coastal HWM locations, across all 11 US storms for which HWMs are available. Further, we do not consider the locations that were classified as dry by the model in the aggregate statistics (Fig. 9a). This means that a different set and number of HWM locations goes into the evaluation for each of the models, ranging from $N = 217$ (GTSM + Aqueduct) to $N = 288$ (GeoClaw) (Fig. 9b). We found that the results do not change significantly when the inundation heights above ground are compared (Supplementary Fig. S3b–d), or when including locations modeled as dry (Supplementary Fig. S3a–c). Finally, we also applied the comparison after restricting to HWMs for which the deviation of the DEM to the recorded elevation does not exceed half the recorded inundation height (Supplementary Fig. S3d; this excludes 90% of the HWMs from the comparison).

We report five different performance scores for the evaluation of flood extents, three of which are overall measures of performance (Fig. 5a–c). In flood extent validation studies, F1 and F2 are the predominant indicators of model performance, even though they are known to be biased in favor of overpredictions[64]. MCC is generally regarded as a more informative and true score if the class sizes vary[65]. Also, contrary to MCC, F1 and F2 are not symmetric, i.e., they will change when exchanging the meanings of "wet" and "dry". We decided to complement the three overall scores by TNR and bias because they represent aspects of the classification problem that are least reflected in F1 and F2, and bias is only implicitly included in MCC.

The performance scores of all models considered in this study appear to be quite low and need to be interpreted in context. They cannot easily be compared to regional analyses or to studies about freshwater or compound flooding. For example, all four approaches considered in this study have average F2 scores of <25%, while F2 scores of <30% are very uncommon in the literature[66,67] (typical values are between 30% and 50%[68,69], even exceeding 80% in some cases[70,71]). In assessments of approaches that model freshwater flooding over land, it is much easier to restrict to a model area that clearly excludes flooding from surge, e.g. for Hurricane Harvey[67]. In contrast, our study is about approaches that model TC-related flooding from storm surge, isolated from freshwater flooding and compound effects. For validation purposes, it would be desirable to have observational flood extents that are purely caused by storm surge. However, in the TC context, storm surge is always accompanied by heavy rainfall so that none of the observational flood extents can clearly be attributed to surge alone. When comparing the surge model outputs with the compound flood extents, low-performance scores are to be expected. There is another aspect that influences the scores seen in studies about freshwater flooding: it is comparably easy for a flood model to correctly classify the permanent water bodies (rivers, lakes) as wet. Still, studies about freshwater flooding usually do not exclude the permanent water bodies from the scoring[68,72]. In our study, it is comparably easy for a surge model to correctly classify the ocean as wet. However, we do not include permanent water bodies in the calculation of the performance scores, since otherwise, the size of the ocean area included in the rectangular flood map would dominate the results. If a flood map containing a larger portion of the ocean scored automatically better than a flood map that only contains a small ocean strip, summarizing scores over a global set of flood maps would be difficult. So far, our study is the first to evaluate a pure TC storm surge modeling approach using a global set of observational flood extents. Therefore, we do not find comparable score values in the literature.

The total scores of GeoClaw are much better than the map-by-map scores. However, the qualitative differences between the models remain mostly unchanged (Fig. 5). Similarly, the qualitative statements are robust to the exclusion of very small and very large (e.g., <100 km² or more than 1000 km²) flood extents from the analysis, when removing the minimum flood threshold of 10 cm, or when considering the flood maps from each data source separately (Supplementary Fig. S1).

Limitations in the considered drivers of flooding

The GTSM data we use are only driven by meteorological forcing from ERA5, while GeoClaw is forced by parametric TC wind fields. Forcing ocean models with ERA5 data is known to underestimate TC-induced surge

heights, and the forcing can be improved by overlaying parametric TC wind fields[4,19,39,79]. On the other hand, the choice of parametric wind field is important in TC storm surge modeling[116], and post-processing methods have been proposed to improve the outputs of parametric models[134].

Coastal flood protection is not included in this analysis, and many exposed regions are protected by dikes and storm surge barriers up to a certain design standard. Many cities have a protection standard that equals a 100-year return period[135], and will be protected against many storms considered in this study. This might be one reason why the tendency of GeoClaw to underestimate hydrographs (Fig. 10a) does not translate to the bias in the flood extent evaluation (Fig. 5e).

In addition to the inclusion of flow dynamics, the different assumptions on friction might explain some of the differences between the dynamic model GeoClaw and the static model Aqueduct (Geoclaw + Aqueduct). Both implement a concept of bottom friction over land. While the "resistance factor" in Aqueduct is conceptually different from the Manning formulation in GeoClaw, both concepts come with a calibration parameter expressing terrain roughness that can have an important impact on overland flow[136].

Note that, while GeoClaw has been used for TC storm surge modeling before[54,137], the main application of GeoClaw is tsunami modeling where it is popular in probabilistic applications with large ensembles of inundation scenarios[138]. Since TC analyses are also often based on probabilistic ensembles of events[9,77,139], this would be an interesting future application of our approach. As an alternative, the one-step approach could in principle also be implemented with more advanced ocean models like ADCIRC, GTSM or SCHISM that are able to account for more complex environmental forcing than GeoClaw[16].

Data availability
All data needed to reproduce the findings reported in this article are openly accessible. A detailed description of the exact versions of the datasets used, including instructions how to download and preprocess these, is provided in the "README.md" file of the archive "evaluation.zip" in the project's Zenodo repository https://doi.org/10.5281/zenodo.10419306. Further, machine-readable source data for all figures, Supplementary Figs., and Supplementary Tables are provided with this paper as Supplementary Data file.

Code availability
All code that was used (i) for the GeoClaw setup, (ii) for the static inundation models, (iii) for the preprocessing of validational data sources, and (iv) to analyze the data and produce the figures was implemented in Python 3.9 (https://www.python.org/) with CLIMADA 3.3.3 (https://zenodo.org/record/7691855), and is openly available from https://doi.org/10.5281/zenodo.10419306.

References
1. CRED/UCLouvain Brussels, B. https://www.emdat.be (EM-DAT, 2022).
2. Zhang, W. et al. Compound hydrometeorological extremes: drivers, mechanisms and methods. *Front. Earth Sci.* **9**, 673495 (2021).
3. Eilander, D. et al. Modeling compound flood risk and risk reduction using a globally applicable framework: a pilot in the Sofala province of Mozambique. *Nat. Hazards Earth Syst. Sci.* **23**, 2251–2272 (2023).
4. Vousdoukas, M. I. et al. Small Island Developing States under threat by rising seas even in a 1.5 °C warming world. *Nat. Sustain.* https://doi.org/10.1038/s41893-023-01230-5 (2023).
5. Mester, B. et al. Human displacements from Tropical Cyclone Idai attributable to climate change. *Nat. Hazards Earth Syst. Sci.* **23**, 3467–3485 (2023).
6. Smiley, K. T. et al. Social inequalities in climate change-attributed impacts of Hurricane Harvey. *Nat. Commun.* **13**, 3418 (2022).
7. Bakkensen, L. A., Park, D.-S. R. & Sarkar, R. S. R. Climate costs of tropical cyclone losses also depend on rain. *Environ. Res. Lett.* **13**, 074034 (2018).
8. Jing, R. et al. Global population profile of tropical cyclone exposure from 2002 to 2019. *Nature* **626**, 549–554 (2024).
9. Geiger, T., Gütschow, J., Bresch, D. N., Emanuel, K. & Frieler, K. Double benefit of limiting global warming for tropical cyclone exposure. *Nat. Clim. Change* **11**, 861–866 (2021).
10. Krichene, H. et al. The social costs of tropical cyclones. *Nat. Commun.* **14**, 7294 (2023).
11. Eberenz, S., Lüthi, S. & Bresch, D. N. Regional tropical cyclone impact functions for globally consistent risk assessments. *Nat. Hazards Earth Syst. Sci.* **21**, 393–415 (2021).
12. Mendelsohn, R., Emanuel, K., Chonabayashi, S. & Bakkensen, L. The impact of climate change on global tropical cyclone damage. *Nat. Clim. Change* **2**, 205–209 (2012).
13. Rappaport, E. N. Fatalities in the United States from Atlantic tropical cyclones: new data and interpretation. *Bull. Am. Meteorol. Soc.* **95**, 341–346 (2014).
14. Islam, M. R., Duc, L., Sawada, Y. & Satoh, M. Does mean sea level trend mask historical storm surge trend: evidence from tropical cyclones affecting Japan since 1980. *Environ. Res. Lett.* **18**, 085004 (2023).
15. Luettich, R. A., Westerink, J. J. & Scheffner, N. W. *ADCIRC: An Advanced Three-Dimensional Circulation Model for Shelves, Coasts, and Estuaries*. Report 1. Theory and Methodology of ADCIRC-2DDI and ADCIRC-3DL. Technical Report DRP-92-6 (U.S. Army Corps of Engineers, Vicksburg, MS, 1992).
16. Loveland, M. et al. Developing a modeling framework to simulate compound flooding: when storm surge interacts with riverine flow. *Front. Clim.* **2**, 609610 (2021).
17. Tsai, Y.-L., Wu, T.-R., Yen, E., Lin, C.-Y. & Lin, S. C. Parallel-computing two-way grid-nested storm surge model with a moving boundary scheme and case study of the 2013 Super Typhoon Haiyan. *Water* **14**, 547 (2022).
18. Kernkamp, H. W. J., Van Dam, A., Stelling, G. S. & de Goede, E. D. Efficient scheme for the shallow water equations on unstructured grids with application to the Continental Shelf. *Ocean Dyn.* **61**, 1175–1188 (2011).
19. Rezaie, A. M. & Haque, A. Development of storm surge inundation model and database for enhanced climate services in Bangladesh. *Front. Water* **4**, 887631 (2022).
20. Chen, C., Liu, H. & Beardsley, R. C. An unstructured grid, finite-volume, three-dimensional, primitive equations ocean model: application to coastal ocean and estuaries. *J. Atmos. Ocean. Technol.* **20**, 159–186 (2003).
21. Weisberg, R. H. & Zheng, L. Hurricane storm surge simulations comparing three-dimensional with two-dimensional formulations based on an Ivan-like storm over the Tampa Bay, Florida region. *J. Geophys. Res. Oceans* **113**, https://doi.org/10.1029/2008JC005115 (2008).
22. Danish Hydraulic Institute (DHI). *MIKE 21 & MIKE 3 Flow Model: Hydrodynamic and Transport Module: Scientific Documentation* (Danish Hydraulic Institute (DHI), Hørsholm, Denmark, 2017).
23. Pan, Z. & Liu, H. Numerical study of typhoon-induced storm surge in the Yangtze Estuary of China using a coupled 3D model. *Procedia Eng.* **116**, 849–854 (2015).
24. Shchepetkin, A. F. & McWilliams, J. C. The regional oceanic modeling system (ROMS): a split-explicit, free-surface, topography-following-coordinate oceanic model. *Ocean Model.* **9**, 347–404 (2005).
25. Qin, G. et al. Storm surge inundation modulated by typhoon intensities and tracks: simulations using the regional ocean modeling system (ROMS). *J. Mar. Sci. Eng.* **11**, 1112 (2023).
26. Jelesnianski, C. P., Chen, J. & Shaffer, W. A. *SLOSH: Sea, Lake, And Overland Surges From Hurricanes*. Technical Report NWS 48 (NOAA, 1992).

27. Zhang, K., Xiao, C. & Shen, J. Comparison of the CEST and SLOSH models for storm surge flooding. *J. Coast. Res.* **24**, 489–499 (2008).

28. Mentaschi, L. et al. A global unstructured, coupled, high-resolution hindcast of waves and storm surge. *Front. Mar. Sci.* **10**, 1233679 (2023).

29. Gori, A., Lin, N., Schenkel, B. & Chavas, D. North Atlantic tropical cyclone size and storm surge reconstructions from 1950-present. *J. Geophys. Res. Atmos.* **128**, e2022JD037312 (2023).

30. Marsooli, R., Lin, N., Emanuel, K. & Feng, K. Climate change exacerbates hurricane flood hazards along US Atlantic and Gulf Coasts in spatially varying patterns. *Nat. Commun.* **10**, 1–9 (2019).

31. Lin, N., Emanuel, K. A., Smith, J. A. & Vanmarcke, E. Risk assessment of hurricane storm surge for New York City. *J. Geophys. Res. Atmos.* **115** https://doi.org/10.1029/2009JD013630 (2010).

32. Stephens, S. A. et al. Future changes in built environment risk to coastal flooding, permanent inundation and coastal erosion hazards. *J. Mar. Sci. Eng.* **9**, 1011 (2021).

33. Ramirez, J. A., Lichter, M., Coulthard, T. J. & Skinner, C. Hyper-resolution mapping of regional storm surge and tide flooding: comparison of static and dynamic models. *Nat. Hazards* **82**, 571–590 (2016).

34. Didier, D. et al. Multihazard simulation for coastal flood mapping: Bathtub versus numerical modelling in an open estuary, Eastern Canada. *J. Flood Risk Manag.* **12**, e12505 (2019).

35. Neumann, T. & Ahrendt, K. *Comparing the "Bathtub Method" with MIKE 21 HD FLOW Model for Modelling Storm Surge Inundation*. Technical Report 22 (Universität Kiel, Kiel, 2013).

36. Vousdoukas, M. I. et al. Developments in large-scale coastal flood hazard mapping. *Nat. Hazards Earth Syst. Sci.* **16**, 1841–1853 (2016).

37. Menéndez, P., Losada, I. J., Torres-Ortega, S., Narayan, S. & Beck, M. W. The global flood protection benefits of mangroves. *Sci. Rep.* **10**, 4404 (2020).

38. Salisbury, M. B. & Hagen, S. C. The effect of tidal inlets on open coast storm surge hydrographs. *Coast. Eng.* **54**, 377–391 (2007).

39. Dullaart, J. C. M. et al. Accounting for tropical cyclones more than doubles the global population exposed to low-probability coastal flooding. *Commun. Earth Environ.* **2**, 1–11 (2021).

40. Kunze, S. & Strobl, E. A. The global long-term effects of storm surge flooding on human settlements in coastal areas. *Environ. Res. Lett.* **19**, 024016 (2024).

41. Kirezci, E. et al. Projections of global-scale extreme sea levels and resulting episodic coastal flooding over the 21st Century. *Sci. Rep.* **10**, 11629 (2020).

42. Ward, P. J. et al. *Aqueduct Floods Methodology*. Technical Note (World Resources Institute, Washington, D.C., 2020).

43. Aznar-Siguan, G. & Bresch, D. N. CLIMADA v1: a global weather and climate risk assessment platform. *Geosci. Model Dev.* **12**, 3085–3097 (2019).

44. Rana, A., Zhu, Q., Detken, A., Whalley, K. & Castet, C. Strengthening climate-resilient development and transformation in Viet Nam. *Clim. Change* **170**, 4 (2022).

45. Dullaart, J. C. M. et al. Enabling dynamic modelling of coastal flooding by defining storm tide hydrographs. *Nat. Hazards Earth Syst. Sci.* **23**, 1847–1862 (2023).

46. MacPherson, L. R., Arns, A., Dangendorf, S., Vafeidis, A. T. & Jensen, J. A stochastic extreme sea level model for the German Baltic Sea coast. *J. Geophys. Res. Oceans* **124**, 2054–2071 (2019).

47. Ayyad, M., Hajj, M. R. & Marsooli, R. Machine learning-based assessment of storm surge in the New York metropolitan area. *Sci. Rep.* **12**, 19215 (2022).

48. Lockwood, J. W., Lin, N., Oppenheimer, M. & Lai, C.-Y. Using neural networks to predict hurricane storm surge and to assess the sensitivity of surge to storm characteristics. *J. Geophys. Res. Atmos.* **127**, e2022JD037617 (2022).

49. Lee, J.-W., Irish, J. L., Bensi, M. T. & Marcy, D. C. Rapid prediction of peak storm surge from tropical cyclone track time series using machine learning. *Coast. Eng.* **170**, 104024 (2021).

50. Liu, Y., Zhao, Q., Hu, C. & Luo, N. Prediction of storm surge water level based on machine learning methods. *Atmosphere* **14**, 1568 (2023).

51. Pachev, B., Arora, P., del-Castillo-Negrete, C., Valseth, E. & Dawson, C. A framework for flexible peak storm surge prediction. *Coast. Eng.* **186**, 104406 (2023).

52. Pringle, W. J., Burnett, Z., Sargsyan, K., Moghimi, S. & Myers, E. Efficient probabilistic prediction and uncertainty quantification of tropical cyclone–driven storm tides and inundation. *Artif. Intell. Earth Syst.* **2**, e220040 (2023).

53. Berger, M. J., George, D. L., LeVeque, R. J. & Mandli, K. T. The GeoClaw software for depth-averaged flows with adaptive refinement. *Adv. Water Resour.* **34**, 1195–1206 (2011).

54. Mandli, K. T. & Dawson, C. N. Adaptive mesh refinement for storm surge. *Ocean Model.* **75**, 36–50 (2014).

55. Muis, S. et al. A high-resolution global dataset of extreme sea levels, tides, and storm surges, including future projections. *Front. Mar. Sci.* **7**, 263 (2020).

56. Murphy, J. D. *Service Assessment: August–September 2017 Hurricane Harvey*. Technical Report (National Weather Service, Silver Spring, Maryland, 2018).

57. Knabb, R. D., Brown, D. P. & Rhome, J. R. *Tropical Cyclone Report: Hurricane Rita*. Technical Report (National Hurricane Center, 2006).

58. Shen, X., Anagnostou, E. N., Allen, G. H., Robert Brakenridge, G. & Kettner, A. J. Near-real-time non-obstructed flood inundation mapping using synthetic aperture radar. *Remote Sens. Environ.* **221**, 302–315 (2019).

59. Brakenridge, G. *Global Active Archive of Large Flood Events* https://floodobservatory.colorado.edu/ (2023).

60. Tellman, B. et al. Satellite imaging reveals increased proportion of population exposed to floods. *Nature* **596**, 80–86 (2021).

61. Bennett, N. D. et al. Characterising performance of environmental models. *Environ. Model. Softw.* **40**, 1–20 (2013).

62. Cea, L., Álvarez, M. & Puertas, J. Estimation of flood-exposed population in data-scarce regions combining satellite imagery and high resolution hydrological-hydraulic modelling: a case study in the Licungo basin (Mozambique). *J. Hydrol. Reg. Stud.* **44**, 101247 (2022).

63. Aronica, G., Bates, P. D. & Horritt, M. S. Assessing the uncertainty in distributed model predictions using observed binary pattern information within GLUE. *Hydrol. Process.* **16**, 2001–2016 (2002).

64. Stephens, E., Schumann, G. & Bates, P. Problems with binary pattern measures for flood model evaluation. *Hydrol. Process.* **28**, 4928–4937 (2014).

65. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6 (2020).

66. Hoch, J. M., Eilander, D., Ikeuchi, H., Baart, F. & Winsemius, H. C. Evaluating the impact of model complexity on flood wave propagation and inundation extent with a hydrologic–hydrodynamic model coupling framework. *Nat. Hazards Earth Syst. Sci.* **19**, 1723–1735 (2019).

67. Chen, M. et al. A comprehensive flood inundation mapping for Hurricane Harvey using an integrated hydrological and hydraulic model. *J. Hydrometeorol.* **22**, 1713–1726 (2021).

68. Willis, T. D. M. et al. Hydrodynamic modeling of inundation patterns of a large African floodplain indicates sensitivity to waterway restoration. *Water Resour. Res.* **58**, e2021WR030107 (2022).

69. Dasgupta, A., Grimaldi, S., Ramsankaran, R. A. A. J., Pauwels, V. R. N. & Walker, J. P. A simple framework for calibrating hydraulic flood inundation models using Crowd-sourced water levels. *J. Hydrol.* **614**, 128467 (2022).

70. Wing, O. E. J. et al. Validation of a 30 m resolution flood hazard model of the conterminous United States. *Water Resour. Res.* **53**, 7968–7986 (2017).

71. Costabile, P., Costanzo, C., Ferraro, D., Macchione, F. & Petaccia, G. Performances of the new HEC-RAS version 5 for 2-D hydrodynamic-based rainfall-runoff simulations at basin scale: comparison with a state-of-the-art model. *Water* **12**, 2326 (2020).

72. Mester, B., Willner, S. N., Frieler, K. & Schewe, J. Evaluation of river flood extent simulated with multiple global hydrological models and climate forcings. *Environ. Res. Lett.* **16**, 094010 (2021).

73. Bates, P. Fundamental limits to flood inundation modelling. *Nat. Water* **1**, 566–567 (2023).

74. USGS. *Short-Term Network (STN) Flood Event Data* https://stn.wim.usgs.gov/ (2023).

75. Haigh, I. D. et al. GESLA version 3: a major update to the global higher-frequency sea-level dataset. *Geosci. Data J.* https://doi.org/10.1002/gdj3.1742 (2022).

76. Knapp, K. R., Kruk, M. C., Levinson, D. H., Diamond, H. J. & Neumann, C. J. The International Best Track Archive for Climate Stewardship (IBTrACS). *Bull. Am. Meteorol. Soc.* **91**, 363–376 (2010).

77. Bloemendaal, N. et al. Global modeling of tropical cyclone storm surges using high-resolution forecasts. *Clim. Dyn.* **52**, 5031–5044 (2019).

78. Muis, S. et al. Spatiotemporal patterns of extreme sea levels along the western North-Atlantic coasts. *Sci. Rep.* **9**, 3391 (2019).

79. Yu, Y.-C. et al. Assessing the potential highest storm tide hazard in Taiwan based on 40-year historical typhoon surge hindcasting. *Atmosphere* **10**, 346 (2019).

80. Mechler, R., Bouwer, L. M., Schinko, T., Surminski, S. & Linnerooth-Bayer, J. (eds) *Loss and Damage from Climate Change: Concepts, Methods and Policy Options*. Climate Risk Management, Policy and Governance (Springer International Publishing, Cham, 2019).

81. Garschagen, M., Doshi, D., Moure, M., James, H. & Shekhar, H. The consideration of future risk trends in national adaptation planning: conceptual gaps and empirical lessons. *Clim. Risk Manag.* **34**, 100357 (2021).

82. Knutson, T. et al. Tropical cyclones and climate change assessment: Part II: projected response to anthropogenic warming. *Bull. Am. Meteorol. Soc.* **101**, E303–E322 (2020).

83. IPCC. *Climate Change 2021: The Physical Science Basis: Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2021).

84. Huang, W. et al. Tidal simulation revisited. *Ocean Dyn.* **72**, 187–205 (2022).

85. Treu, S. et al. Reconstruction of hourly coastal water levels and counterfactuals without sea level rise for impact attribution. *Earth Syst. Sci. Data* **16**, 1121–1136 (2024).

86. Lee, W., Sun, A. Y., Scanlon, B. R. & Dawson, C. Hindcasting compound pluvial, fluvial and coastal flooding during Hurricane Harvey (2017) using Delft3D-FM. *Nat. Hazards* https://doi.org/10.1007/s11069-023-06247-9 (2023).

87. Valle-Levinson, A., Olabarrieta, M. & Heilman, L. Compound flooding in Houston-Galveston Bay during Hurricane Harvey. *Sci. Total Environ.* **747**, 141272 (2020).

88. Leal Filho, W. & Nalau, J. (eds.) *Limits to Climate Change Adaptation*. Climate Change Management (Springer International Publishing, Cham, 2018).

89. Kerns, B. W. & Chen, S. S. Compound effects of rain, storm surge, and river discharge on coastal flooding during Hurricane Irene and Tropical Storm Lee (2011) in the Mid-Atlantic region: coupled atmosphere-wave-ocean model simulation and observations. *Nat. Hazards* **116**, 693–726 (2023).

90. Zeng, Z., Xu, J., Ye, G. & Shen, W. The influence of different intensity of monsoon on typhoon precipitation: a comparative study of typhoons Soudelor and Maria. *Front. Earth Sci.* **11**, 1251711 (2023).

91. Xi, D., Lin, N. & Gori, A. Increasing sequential tropical cyclone hazards along the US East and Gulf coasts. *Nat. Clim. Change* **13**, 258–265 (2023).

92. Fu, Z.-H., Zhan, R., Zhao, J., Yamada, Y. & Song, K. Future projections of multiple tropical cyclone events in the Northern Hemisphere in the CMIP6-HighResMIP models. *Geophys. Res. Lett.* **50**, e2023GL103064 (2023).

93. Tarpanelli, A., Mondini, A. C. & Camici, S. Effectiveness of Sentinel-1 and Sentinel-2 for flood detection assessment in Europe. *Nat. Hazards Earth Syst. Sci.* **22**, 2473–2489 (2022).

94. Grimaldi, S., Li, Y., Pauwels, V. R. N. & Walker, J. P. Remote sensing-derived water extent and level to constrain hydraulic flood forecasting models: opportunities and challenges. *Surv. Geophys.* **37**, 977–1034 (2016).

95. Brivio, P. A., Colombo, R., Maggi, M. & Tomasoni, R. Integration of remote sensing data and GIS for accurate mapping of flooded areas. *Int. J. Remote Sens.* **23**, 429–441 (2002).

96. Wijetunge, J. J. & Neluwala, N. G. P. B. Compound flood hazard assessment and analysis due to tropical cyclone-induced storm surges, waves and precipitation: a case study for coastal lowlands of Kelani river basin in Sri Lanka. *Nat. Hazards* **116**, 3979–4007 (2023).

97. Liang, H. & Zhou, X. Impact of tides and surges on fluvial floods in coastal regions. *Remote Sens.* **14**, 5779 (2022).

98. Bates, P. D. et al. Combined modeling of US fluvial, pluvial, and coastal flood hazard under current and future climates. *Water Resour. Res.* **57**, e2020WR028673 (2021).

99. GEBCO. *General Bathymetric Chart of the Oceans (GEBCO)* https://www.gebco.net/ (2014).

100. Muis, S. et al. Global projections of storm surges using high-resolution CMIP6 climate models. *Earth's Future* **11**, e2023EF003479 (2023).

101. Tozer, B. et al. Global bathymetry and topography at 15 arc sec: SRTM15+. *Earth Space Sci.* **6**, 1847–1864 (2019).

102. Kulp, S. A. & Strauss, B. H. *CoastalDEM v2.1: A High-accuracy and High-resolution Global Coastal Elevation Model Trained on ICESat-2 Satellite Lidar*. Technical Report (Climate Central Inc., Princeton, USA, 2021).

103. Yamazaki, D. et al. MERIT Hydro: a high-resolution global hydrography map based on latest topography dataset. *Water Resour. Res.* **55**, 5053–5073 (2019).

104. Acosta-Morel, M., McNulty, V. P., Lummen, N., Schill, S. R. & Beck, M. W. Shoreline solutions: guiding efficient data selection for coastal risk modeling and the design of adaptation interventions. *Water* **13**, 875 (2021).

105. Meadows, M., Jones, S. & Reinke, K. Vertical accuracy assessment of freely available global DEMs (FABDEM, Copernicus DEM, NASADEM, AW3D30 and SRTM) in flood-prone environments. *Int. J. Digit. Earth* **17**, 2308734 (2024).

106. Lyard, F. H., Allain, D. J., Cancet, M., Carrère, L. & Picot, N. FES2014 global ocean tide atlas: design and performance. *Ocean Sci.* **17**, 615–649 (2021).

107. Copernicus Marine Service. *Global Ocean Gridded L4 Sea Surface Heights And Derived Variables Reprocessed (1993-Ongoing)* https://doi.org/10.48670/MOI-00148 (2021).

108. Holland, G. J. An analytic model of the wind and pressure profiles in hurricanes. *Mon. Weather Rev.* **108**, 1212–1218 (1980).

109. Frieler, K. et al. Scenario setup and forcing data for impact model evaluation and impact attribution within the third round of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP3a). *Geosci. Model. Dev.* **17**, 1–51 (2024).

110. Garratt, J. R. Review of drag coefficients over oceans and continents. *Mon. Weather Rev.* **105**, 915–929 (1977).

Article

111. Charnock, H. Wind stress on a water surface. *Q. J. R. Meteorol. Soc.* **81**, 639–640 (1955).

112. Sampson, C. C. et al. A high-resolution global flood hazard model. *Water Resour. Res.* **51**, 7358–7381 (2015).

113. Qin, X., LeVeque, R. J. & Motley, M. R. Accelerating an adaptive mesh refinement code for depth-averaged flows using GPUs. *J. Adv. Model. Earth Syst.* **11**, 2606–2628 (2019).

114. Du, H., Yu, P., Zhu, L., Fei, K. & Gao, L. Assessing the performances of parametric wind models in predicting storm surges in the Pearl River Estuary. *J. Wind Eng. Ind. Aerodyn.* **232**, 105265 (2023).

115. Wang, S., Lin, N. & Gori, A. Investigation of tropical cyclone wind models with application to storm tide simulations. *J. Geophys. Res. Atmos.* **127**, e2021JD036359 (2022).

116. Yan, D. & Zhang, T. Research progress on tropical cyclone parametric wind field models and their application. *Reg. Stud. Mar. Sci.* **51**, 102207 (2022).

117. Holland, G. J., Belanger, J. I. & Fritz, A. A revised model for radial profiles of hurricane winds https://doi.org/10.1175/2010MWR3317.1 (2010).

118. Vijayan, L. et al. Evaluation of parametric wind models for more accurate modeling of storm surge: a case study of Hurricane Michael. *Nat. Hazards* **106**, 2003–2024 (2021).

119. Irazoqui Apecechea, M., Verlaan, M., Zijl, F., Le Coz, C. & Kernkamp, H. Effects of self-attraction and loading at a regional scale: a test case for the Northwest European Shelf. *Ocean Dyn.* **67**, 729–749 (2017).

120. Hersbach, H. et al. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **146**, 1999–2049 (2020).

121. Holland, G. J. A revised hurricane pressure-wind model. *Mon. Weather Rev.* **136**, 3432–3445 (2008).

122. Tiggeloven, T. et al. Global-scale benefit–cost analysis of coastal flood adaptation to different flood risk drivers using structural measures. *Nat. Hazards Earth Syst. Sci.* **20**, 1025–1044 (2020).

123. Pekel, J.-F., Cottam, A., Gorelick, N. & Belward, A. S. High-resolution mapping of global surface water and its long-term changes. *Nature* **540**, 418–422 (2016).

124. Haer, T. et al. Coastal and river flood risk analyses for guiding economically optimal flood adaptation policies: a country-scale study for Mexico. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **376**, 20170329 (2018).

125. Samela, C., Coluzzi, R., Imbrenda, V., Manfreda, S. & Lanfredi, M. Satellite flood detection integrating hydrogeomorphic and spectral indices. *GIScience Remote Sens.* **59**, 1997–2018 (2022).

126. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta (BBA) Protein Struct.* **405**, 442–451 (1975).

127. Stanski, H. R., Wilson, L. J. & Burrows, W. R. *Survey of Common Verification Methods in Meteorology*. Research Report MSRB 89-5 (Atmospheric Environment Service, Forecast Research Division, Canada, 1989).

128. Bernhofen, M. V. et al. A first collective validation of global fluvial flood models for major floods in Nigeria and Mozambique. *Environ. Res. Lett.* **13**, 104007 (2018).

129. Bhattacharya, B., Mazzoleni, M. & Ugay, R. Flood inundation mapping of the sparsely gauged large-scale Brahmaputra basin using remote sensing products. *Remote Sens.* **11**, 501 (2019).

130. Koenig, T. A. et al. Identifying and preserving high-water mark data. in *U.S. Geological Survey Techniques and Methods*, 47 (U.S. Geological Survey, Reston, Virginia, 2016).

131. Yang, K., Paramygin, V. A. & Sheng, Y. P. A rapid forecasting and mapping system of storm surge and coastal flooding. *Weather Forecast.* **35**, 1663–1681 (2020).

132. Gutenson, J. L. et al. Comparison of estimated flood exposure and consequences generated by different event-based inland flood inundation maps. *Nat. Hazards Earth Syst. Sci.* **23**, 261–277 (2023).

133. Ferguson, S. et al. Assessing numerical model skill at simulating coastal flooding using field observations of deposited debris and photographic evidence. *Water* **14**, 589 (2022).

134. Iwamoto, T., Takagawa, T., Shibayama, T., Esteban, M. & Mäll, M. A proposal of a semi-empirical method for modifying the atmospheric pressure and wind fields of tropical cyclones. *Coast. Eng. J.* **65**, 418–432 (2023).

135. Hallegatte, S., Green, C., Nicholls, R. J. & Corfee-Morlot, J. Future flood losses in major coastal cities. *Nat. Clim. Change* **3**, 802–806 (2013).

136. Vafeidis, A. T. et al. Water-level attenuation in global-scale assessments of exposure to coastal flooding: a sensitivity analysis. *Nat. Hazards Earth Syst. Sci.* **19**, 973–984 (2019).

137. Toyoda, M., Fukui, N., Miyashita, T., Shimura, T. & Mori, N. Uncertainty of storm surge forecast using integrated atmospheric and storm surge model: a case study on Typhoon Haishen 2020. *Coast. Eng. J.* **64**, 135–150 (2022).

138. Salazar-Monroy, E. F., Melgar, D., Jaimes, M. A. & Ramirez-Guzman, L. Regional probabilistic tsunami hazard analysis for the Mexican subduction zone from stochastic slip models. *J. Geophys. Res. Solid Earth* **126**, e2020JB020781 (2021).

139. Meiler, S. et al. Intercomparison of regional loss estimates from global synthetic tropical cyclone models. *Nat. Commun.* **13**, 6156 (2022).

Author contributions
T.V., C.O., and K.F. designed the analysis. T.V. implemented the inundation models and calculated the flood depth maps with the support of S.T. T.V. further conducted the statistical analysis, and generated all plots and tables. All authors contributed to the analysis. T.V., C.O., and M.M. wrote the manuscript with contributions from all authors. All authors discussed the results.

Funding
Open Access funding enabled and organized by Projekt DEAL.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43247-024-01707-x.

Correspondence and requests for materials should be addressed to Christian Otto.

Peer review information *Communications Earth & Environment* thanks Kees Nederhoff and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Olusegun Dada and Joe Aslin. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.