

Storm surge forecasts from a global tide-gauge network

Heshan Chen

heshan@stanford.edu

Stanford University <https://orcid.org/0009-0006-0388-0279>

andrew kruczkiewicz

Article

Keywords:

Posted Date: September 25th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11077121/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

Storm surge forecasts from a global tide-gauge network

Heshan Chen^{1,2}, Andrew Kruczkiewicz¹

¹Columbia Climate School, Columbia University, New York, NY, USA. ²Present address: Department of Civil and Environmental Engineering, Stanford University, Stanford, CA, USA. Correspondence: heshan@stanford.edu

Abstract. Storm surge is the deadliest component of coastal flooding, yet access to local surge forecasts remains uneven. Here we present a model, trained on 760 tide gauges, that forecasts surge at gauges and regions it was never trained on from about nine days of local water level and open weather data. It beats persistence at 81 of 84 test gauges (pooled skill +37%, storm windows +32 to +40%) and in four regions withheld from training (+8 to +30%). The gains persist under real weather forecasts (+30% over two years). Transfer depends on the diversity of the training gauges; skill grows with their number. Given the same local observation it beats the global hydrodynamic reanalysis at an eight-hour lead and matches it in storms; its 99th-percentile envelope covers under half the largest peaks. Gauges on coast types the network represents poorly would add more skill than gauges where it is dense.

Storm surge is the deadliest component of coastal flooding; it accounts for roughly half of tropical-cyclone fatalities in the United States¹, and without further adaptation 0.2 to 4.6% of the world’s population is projected to be flooded annually by 2100, with expected annual losses of up to 9.3% of global GDP² and up to US\$14.2 trillion in exposed assets³. However, forecast quality is unevenly distributed, with a third of the world’s population, mostly in least developed countries and small island states, not covered by any early warning system⁴. Climate change is projected to lead to higher sea levels while socioeconomic changes, some of which may be influenced by climate change, are likely to increase the number of coastal populations exposed to storm surge extreme events, which is one component of a likely shift in distribution of storm surge and compound event risk^{3,5,6}.

While early warning systems (EWS) for storm surge exist, these are primarily functional in countries with dedicated national-level sustained funding^{7,8} or where there is a longstanding, structured agreement with a regional entity or other country providing funding for production, implementation, monitoring and evaluation^{9,10}. However, in each context, the recent shifts in government priorities and geopolitical agreements have led to an urgency to review current governance structures for storm surge forecasting, and explore new approaches^{11–13}.

Further, with the UN Secretary-General’s flagship programme Early Warnings for All, there is growing interest in implementing warning systems for all hazards for the countries¹⁴, and districts of countries¹⁵, that are traditionally underserved in that regard^{16–18}.

Numerical surge models (ADCIRC, GTSM¹⁹) provide the operational backbone where they exist. Their coverage, latency, and computational cost leave most of the world’s coastline without skilful local surge guidance. Machine learning has addressed parts of this challenge: models trained separately at each gauge (station-specific models) match physics-based models at that gauge^{20,21}, station-specific statistical models cover the

global gauge network at daily-maximum resolution²², and 2024 to 2026 produced a wave of attempts to generalize beyond the training gauges^{23–26}, alongside regional emulators conditioned on station attributes, limited-area surge models, and post-processors of simulation output^{27–29}.

Each attempt stops short of the deployment question. Station-specific models, including the global daily-maximum reconstructions of Tadesse et al.²², cannot predict at sites without local training data²⁰. A pooled model trained on more than a thousand gauges between 45°S and 45°N reconstructs historical surge at unmonitored sites³⁰; it is a reanalysis-driven hindcast, validated by random station folds, with no forecast lead time and no test on a withheld coastline. Regional transfer stays within one coastline: 31 Florida gauges²⁴; 14 US Atlantic targets²⁵, where reported skill depends on model selection using target-station observations, which no ungauged coast can execute. Most global machine-learning surge models are trained on synthetic simulation output and evaluated on random storm splits²⁶, are trained on a surge reanalysis and scored in basins that hold few training storms³¹, or require registered simulation products as dense inputs²³. No published work trains on tide-gauge observations at global scale and then tests the model on entire regions withheld from training; none reports how skill scales with the size and composition of the gauge network; and, apart from ensembles calibrated and scored at their own training gauges^{20,32}, none produces calibrated probabilistic forecasts at coasts the model was not trained on, even though what a warning service needs during a storm is the probability that the water level will exceed a threshold³³. Transfer, the ability of a model trained at some gauges to forecast at others, is the standard test for prediction at ungauged sites in hydrology^{34,35} and the route proposed for flood prediction at data-limited coasts³⁶. This leads to two testable hypotheses: that a model trained on the global gauge network transfers to coasts outside it, and that the transfer depends on the diversity of the training gauges.

Here we test these hypotheses with a global dataset of 1,054 quality-screened tide gauges with open reanalysis forcing (Fig. 1a) and one recurrent neural network (Fig. 1b), trained once and then evaluated only at gauges and whole regions that contributed no training data. We evaluate the model under reanalysis forcing and under real weather forecasts, measure what a local water-level feed adds, compare it with a global hydrodynamic model given the same local information, and report where its probabilistic envelope fails. The dataset recipe, splits, protocol, and models are released.

Results

Transfer to unseen gauges on every continent

On the pre-registered marine test set (84 open-coast gauges on every inhabited continent) the delivered neural network (hereafter the model), trained on 760 gauges under the clean protocol (Methods), forecasts zero-shot, with no training data from any test gauge, and beats persistence at 81 of 84 gauges: pooled skill +37% (region-block bootstrap interval +33 to +41%), 8-h skill +30% (+26 to +34%), NNSE at 8 h 0.782 (Table 1, Fig. 2; metric definitions in Supplementary Table 1). Skill is positive at every lead from 1 to 48 h (+11% at 1 h, rising to about +40% at 48 h), and skill inside storm windows is higher still at every storm severity level, from the 90th to the 99.9th percentile of surge (Fig. 2b). Every continent is positive, including the two with the fewest training gauges (Africa +31%, South America +37%). The three gauges where the model does not beat persistence are Dakar (Senegal), Veracruz (Mexico), and Varennes (Canada), the last on the St Lawrence River upstream of Quebec City (Supplementary Fig. 1a). Skill during extreme conditions is higher than skill over all hours. Restricted to hours when observed surge exceeds a gauge’s 99th to 99.9th percentile (high-water hours), skill is +45 to +48%; within storm windows, pooled over all storm windows, it is +32 to +40% at 4 to 12 h lead and about +47% at 48 h; and at the hour of each storm’s peak the forecast reaches on average 59% of the observed peak height, against 24% for persistence (peak capture, Methods). The six largest test-gauge storms, forecast zero-shot, are shown in Extended Data Fig. 1, their storm context in Extended Data Fig. 2, and twelve unselected forecast windows in Supplementary Fig. 2. The q99 envelope’s hourly coverage is 0.98 to 0.99 at every lead, yet it lies above the true peak in only 41% of p99.9 windows (interval 31 to 51%): its mean height at the peak hour is 0.91 to 0.99 of the mean peak height, a ratio that averages over under- and over-prediction. All primary numbers are computed on the 71% of evaluation windows that are 256 consecutive hours (Methods); on

Table 1: Deployment evaluation: zero-shot skill on the 84 pre-registered marine test gauges (760-gauge training set, validation-fold checkpoint selection, continuous evaluation windows). Skills against persistence; wins = gauges beating persistence on pooled RMSE; block-bootstrap intervals for the pooled rows are given in Methods. Rows group gauges geographically: Africa includes the Canary Islands and Madeira; North America includes Mexico and the US Virgin Islands; South America includes the Falkland Islands and the Antarctic Peninsula station; Pacific islands are the seven island gauges of the tropical Pacific outside Australia.

Region	<i>n</i>	pooled	@8h	NNSE@8h	wins
Asia	18	+39%	+30%	0.790	18/18
Europe	18	+40%	+29%	0.803	18/18
N. America	15	+37%	+29%	0.776	13/15
Oceania	10	+35%	+26%	0.855	10/10
S. America	8	+37%	+37%	0.725	8/8
Africa	8	+31%	+35%	0.762	7/8
Pacific islands	7	+23%	+32%	0.697	7/7
All	84	+37%	+30%	0.782	81/84

all windows pooled skill is +39%, 8-h skill +31%, and 82 of 84 gauges are positive; every headline number is reported on both sets in Methods.

Perfect-prognosis forcing, in which a reanalysis stands in for the forecast atmosphere³⁷, measures what a forecast system inherits from a perfect atmosphere, so we also drove the delivered model with real forecasts: GEFSv12 reforecast control-member forcing at every 00 UTC cycle of 2012 and 2017, at every test gauge with a continuous record (79 gauges, 35,228 forecast windows, 283 of them containing a p99.9 storm), against the same model with reanalysis forcing and against persistence on identical windows (Extended Data Table 1, Extended Data Fig. 3; Methods). With real forcing the model keeps four fifths of its skill: pooled skill +30% against +37% with reanalysis forcing, 8-h skill +24% against +28%, 48-h skill +33% against +41%, and it beats persistence at 72 of 79 gauges (reanalysis forcing 74 of 79); the median per-gauge cost of real forcing is 5% of RMSE, and the two years agree to within a point. Inside storm windows RMSE rises from 13.4 to 15.4 cm against 24.7 cm for persistence, and peak capture and envelope coverage are unchanged within sampling (Extended Data Table 1). For the nine largest test-gauge storm cases with complete coverage (eight storms, Xaver at two gauges), issued 27 to 45 h before the peak, the loss of skill from using real forecasts is larger: storm-window RMSE grows from 37.1 to 49.3 cm, point peak capture falls from 0.47 to 0.36, and 8 of 9 cases still beat persistence (Supplementary Table 2). Real forecast forcing costs least on everyday accuracy and most on the magnitude of the largest peaks at more than a day of lead.

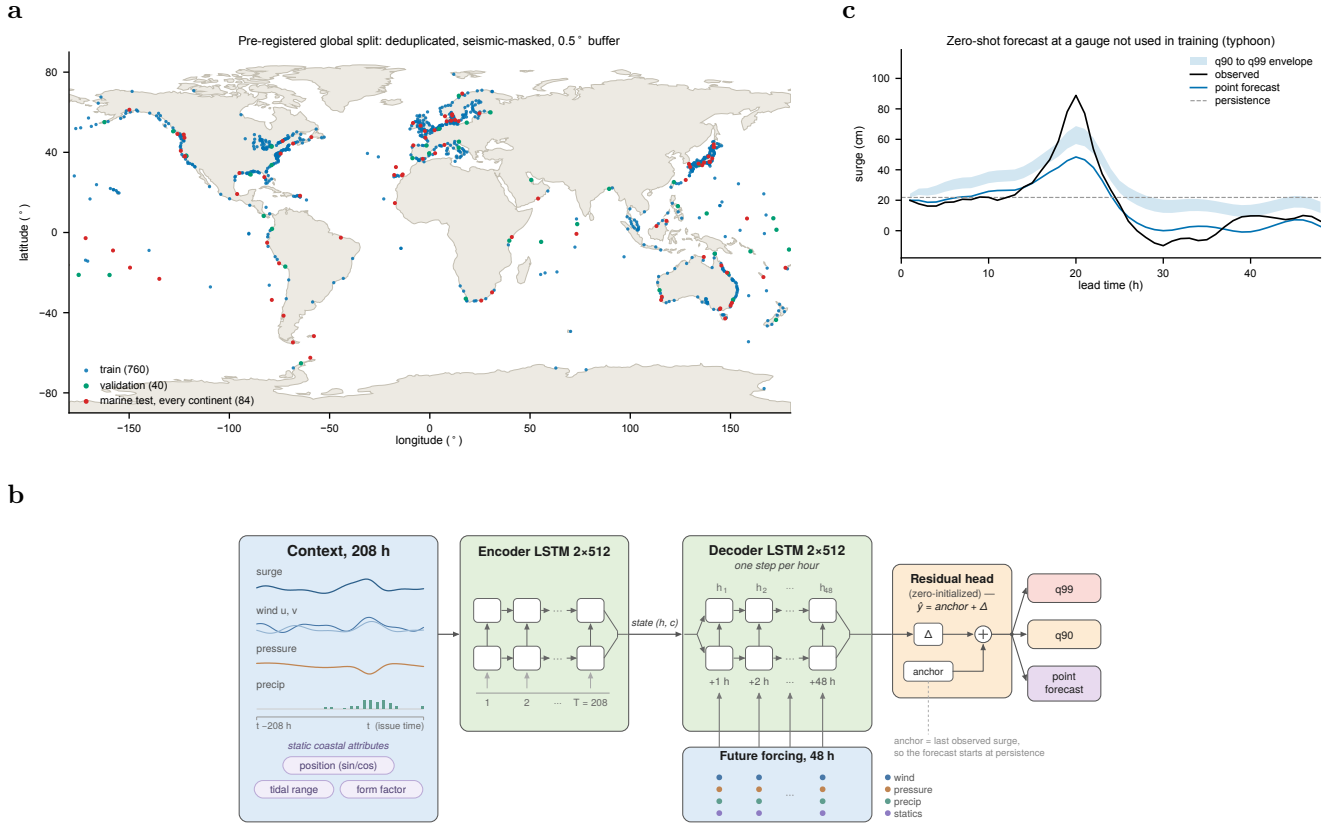


Figure 1: Study design. (a) Global split after QC screening, deduplication, seismic masking, and 0.5° exclusion buffers: training (blue, 760), validation (green, 40), and the primary marine test set (red, 84, every inhabited continent). The 16 non-marine pre-registered test gauges are excluded from the primary set (Methods). (b) The delivered model: a sequence-to-sequence recurrent network whose decoder consumes each hour’s future forcing, with static coastal attributes at every step and a zero-initialized residual head anchored at persistence emitting point, q90, and q99 trajectories. (c) Example output at a test gauge the model was not trained on, during a typhoon (clean-protocol rotation model): peak timing is captured, and the q90 to q99 envelope carries most of the peak height that the point forecast under-predicts.

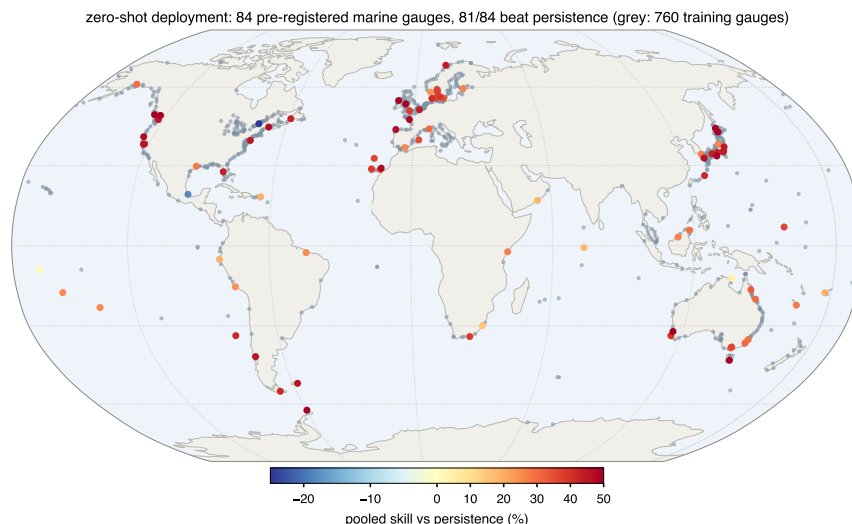
Transfer to whole regions never used in training

To test whether the model can forecast an entire region that contributed no training data, we trained four further models, each with one region removed from the training set (Japan and the Pacific, Europe, North America, Oceania), and evaluated each on the region it had never seen, the approach used to test hydrological models on ungauged basins³⁴. Every withheld region beats persistence on pooled and 8-h skill (Fig. 3; Extended Data Table 2); the only negative value is the low end of Oceania’s storm-window range. The union gives 633 zero-shot gauge evaluations: pooled +11%, 8-h +11%, with 477 of 633 gauges beating persistence (region-block bootstrap intervals for each rotation in Methods). Japan (+30% pooled, storm per-lead +27 to +47%) and North America (+10%, +17 to +20%) transfer strongly on storms; Europe (+11% pooled) has the lowest storm-window skill of the large regions (+8

to +13%), and the ordering matches how well each region’s surge regime, the characteristic response of its coast type to weather, is represented elsewhere in the training set: tropical-cyclone surge occurs on three continents, but the North Sea’s wind-driven surge over a wide shallow shelf has no counterpart outside Europe.

We tested this explanation directly by changing the training set while keeping the European test gauges fixed. Holding the European test set fixed (231 gauges) and training two Europe-free training sets under the same protocol, a 299-gauge training set reaches pooled skill of zero (−0.4%), storm per-lead skill +5 to +8%, and high-water-hour skill +8% (p99); the 593-gauge training set, still with zero European data, reaches +11% pooled (lower error at 186 of 231 gauges; 8-h RMSE difference −0.5 to −0.2 cm), +8 to +13% in storms, and +21% at high water, with peak capture rising from 0.29 to 0.38. Regime transfer can be bought with additional non-local data. This experiment changes the number and the diversity of training

a



b

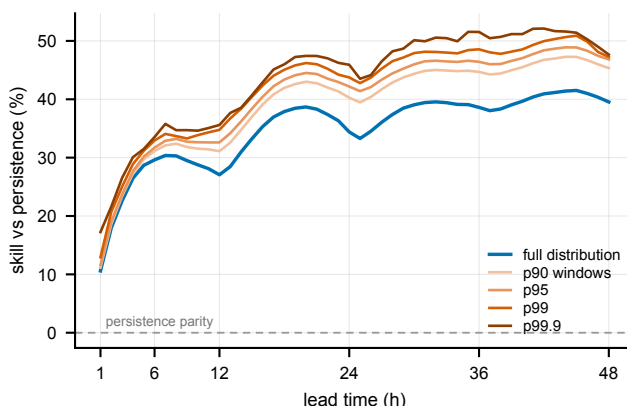


Figure 2: Deployment skill (continuous evaluation windows; Methods). (a) The 84 pre-registered marine test gauges coloured by pooled RMSE skill against persistence (delivered 760-gauge model; positive is better than persistence); grey points are the training network. 81 of 84 gauges are positive, on every inhabited continent. (b) Skill against persistence at every lead from 1 to 48 h, for the full distribution and for storm windows at four severities (p90 to p99.9). Storm-window curves are averages over gauges; the storm skills quoted in the text pool all storm windows across gauges.

gauges at once; the next one changes only the diversity, at a fixed number of gauges.

210 The Oceania experiment shows why checkpoint selec-
tion must never look at the test region. Under check-
point selection on the withheld region’s own skill, Ocea-
nia scored +17% pooled (all windows); under the final
protocol it scores +8% (median of three runs; +11% on
215 all windows), and the gap sits in one basin type: the
enclosed lagoons of Port Phillip Bay and the Gippsland
Lakes are negative in all three clean runs yet positive
under test-side selection, which had picked the one of
four candidate checkpoints that fit this thinly repre-
220 sented coast type (Extended Data Fig. 4). Water levels
inside narrow-mouthed lagoons respond to wind over

the lagoon itself and to the state of the inlet³⁸, which
a model driven by the weather at a single point cannot
see; this limit appears in every run and belongs to the
225 setting, not to a particular model.

Skill scales with gauge count and regime diversity

Seven training-set sizes drawn from the same clean pool
(64 to 760 gauges, country-stratified, composition held
proportional, identical recipe) were evaluated on the
230 fixed 84-gauge test set (Fig. 4; Extended Data Table 3).
Point and extreme-event skill rise across the full range:
pooled skill gains about 5 points and 8-h skill about 4
per doubling of the training set (log-linear fit), and skill

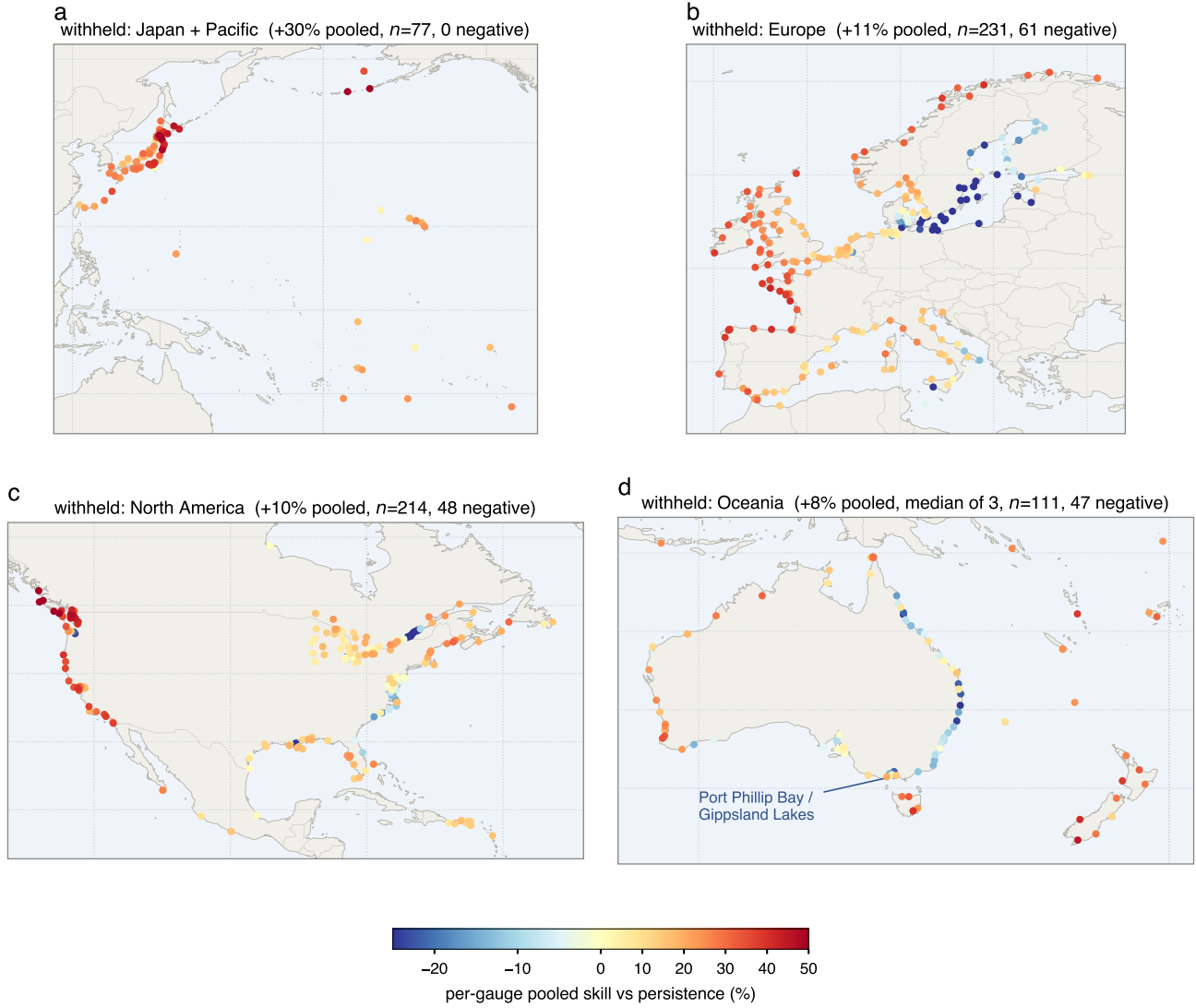


Figure 3: Out-of-network transfer: each panel shows one leave-region-out rotation, every gauge coloured by its zero-shot pooled skill under a model trained without any data from that region (Oceania: median of three runs). Negative counts per rotation are given in the panel titles; the Oceania panel marks the enclosed-lagoon cluster of Extended Data Fig. 4.

was still rising at the largest training-set sizes, though by the smallest amounts (1 point pooled and 0.6 at 8 h from 640 to 760 gauges). Extreme-event skill scales fastest (high-water p99.9 skill +26 to +45%) and calibration improves alongside (Extended Data Table 3). Every cross-size gap in pooled skill exceeds the 0.8-point difference between two same-pipeline runs at 512 gauges (the 640-to-760 step by a small margin), and the same step at 8 h does not (Methods). Every training-set size trains for 12 epochs, so optimizer steps grow about 12-fold from 64 to 760 gauges; retraining the 64- and 256-gauge training sets for the 760-gauge step budget moves them by at most two points (Methods), so twelve times the updates buy at most two points while doubling the gauges buys about five.

Two experiments separate the number of training gauges from their diversity: the European experiment above, and a comparison at a fixed number of gauges in which only the countries they come from differ, evaluated on the same 84 test gauges with the same protocol. Two hundred and ninety-eight gauges country-stratified across 49 countries reach +29% pooled and +24% at 8 h, between the neighbouring 256- and 384-gauge sizes (+28 and +31%); the 298 gauges of the United States and Japan alone reach +7% pooled and -2% at 8 h (the diverse training set has the lower 8-h error at 67 of 84 gauges; difference -2.9 to -0.6 cm). The two-country model keeps its skill where its regimes have analogues (Japan +30% at 8 h, the United States and Canada +20%) and turns negative where they do not (Europe

265 -14% , Australia -70%). The extra storm skill at Eu-
 270 ropean gauges came from the 324 training gauges added
 outside Europe, mostly in North America, Japan, and
 Australia, among them the wide shallow shelves of the
 North American Atlantic and Gulf coasts. A training
 set of the same size drawn from two countries does not
 supply it (Europe -14% at 8 h): what transfers is cov-
 erage of a coast type, which more gauges of the same
 type cannot replace.

The value of local information

275 The delivered model reads 208 h of the target gauge's
 own water level and anchors its forecast on the last
 observation. Retraining the same architecture with
 the water-level input and the anchor removed yields
 a gauge-free variant that does not consume real-time
 280 water level; its outputs are expressed in units of the
 gauge's typical surge size, the standard deviation of its
 surge record (σ_s , the surge scale), and every gauge-
 specific input can be replaced by an openly available
 global product. On the same 84 gauges, the ladder
 285 from no local information to a live gauge is measured
 in Table 2a. With every descriptor from open products
 (tidal range and form factor from the EOT20 global
 tide model³⁹, the surge scale predicted from reanalysis-
 forcing covariates), the variant reaches NNSE at 8 h of
 0.593 (67 of 84 gauges above climatology) with envelope
 290 coverage near nominal and no quantile crossing. Local
 harmonic tidal constants add 0.005, the measured local
 surge scale 0.020, and the real-time water level 0.164 (to
 0.782). Event-conditioned skill survives the removal of
 all local information: inside p99.9 storm windows the
 295 all-open configuration captures 0.50 of true peak height
 with its point forecast (flagship 0.59, persistence 0.24),
 and its q99 envelope at the true peak hour averages 0.99
 of peak height, because storm peaks are largely forcing-
 driven and the response scale is predictable from open
 300 covariates. An inference-time ablation of the gauge-
 fed flagship (zeroing its water-level input) reaches only
 0.522: gauge-free skill must be trained in. The closest
 published result for coasts without gauges, NNSE 0.556
 305 from a simulation-based method that uses ERA5 fields,
 hydrodynamic model output, and neighbouring gauges
 as inputs²³, is of the same order as our all-open vari-
 ant, which uses strictly less information at inference;
 the two studies use different test sets, so the compari-
 310 son is indirect (see Methods).

The comparison supports the importance of a sus-
 tained local water-level observation: the real-time feed
 is the largest single term in the ladder, local calibration
 constants add little on top, and the physics adds to the
 315 observation inside storms (next section).

Comparison with a global hydrodynamic model given the same local observation

A coastal service without a local surge model faces a
 practical question: what does one tide gauge plus an
 open-data model add, relative to the global physics
 product it may or may not be able to access? We
 ran the head-to-head against the ERA5-forced GTSM
 v3 surge reanalysis¹⁹, the only registered product used
 in this study, with the local information on both sides
 325 matched (Table 2b; Fig. 5; Methods): nearest output
 point, the reanalysis high-passed with the same 30-day
 filter as the target (Methods), identical 48-h windows
 over 2010–2018 at 80 gauges, and two tiers of local
 information. The two systems differ in meteorologi-
 330 cal input (ERA5 fields over the whole ocean against
 the reanalysis-forcing series at the gauge), so the head-
 to-head compares two complete forecasting systems at
 equal local observation (Methods). Without any local
 observation (tier A), no difference is detectable between
 the gauge-free model and the reanalysis on the full dis-
 335 tribution (38 of 80 gauges; block-bootstrap 95% inter-
 val of the 8-h RMSE difference -0.5 to $+1.9$ cm; NNSE
 0.60 against 0.62), and the reanalysis is 26% better in-
 side storm windows (8.8 against 11.8 cm at 8 h; the
 intervals at 8 and 48 h exclude zero). With the obser-
 340 vation at issue time on both sides (tier B; the reanalysis
 corrected by its mean error over the preceding 25 h, the
 standard operational adjustment), the neural network
 has the lower full-distribution error at 8 h (66 of 80
 345 gauges; difference -2.1 to -0.2 cm; NNSE 0.80 against
 0.67) and no detectable difference at 48 h (-1.2 to $+0.8$
 cm). Inside storm windows the differences are not sta-
 tistically distinguishable from zero: at 8 h (7.9 against
 7.7 cm) or at 48 h (11.2 against 9.6 cm, interval -0.9 to
 350 $+5.6$ cm), nor in peak capture (0.62 against 0.66, -0.09
 to $+0.01$); the physics model has the better mean value
 in each case, but the intervals include zero, so the data
 cannot tell the two systems apart there; they do not
 show that the two are equal. All intervals come from
 355 resampling regions (Methods).

The map locates the physics advantage (Fig. 5). Out-
 side the Baltic and North Sea the forecast wins bulk ac-
 curacy at 65 of 71 gauges; inside it the reanalysis wins 8
 of 9 (with 22 to 46% lower error) and every peak at the
 five gauges with enough storm windows, because Baltic
 surge is a basin-filling response to wind over the whole
 sea whose information sits in the spatial field the hy-
 360 drodynamic model consumes and single-point forcing
 cannot see. Persistence alone beats the uncorrected,
 mean-aligned reanalysis at 63 of 80 gauges at 8 h, and
 the corrected reanalysis adds nothing to the observation
 on the bulk at that lead (40 of 80; NNSE 0.67 for both)
 while adding much inside storms (7.7 against 12.4 cm).
 The observation carries short-lead bulk skill; without
 370 an observation the physics has the lower storm-window
 error; with the observation on both sides the neural

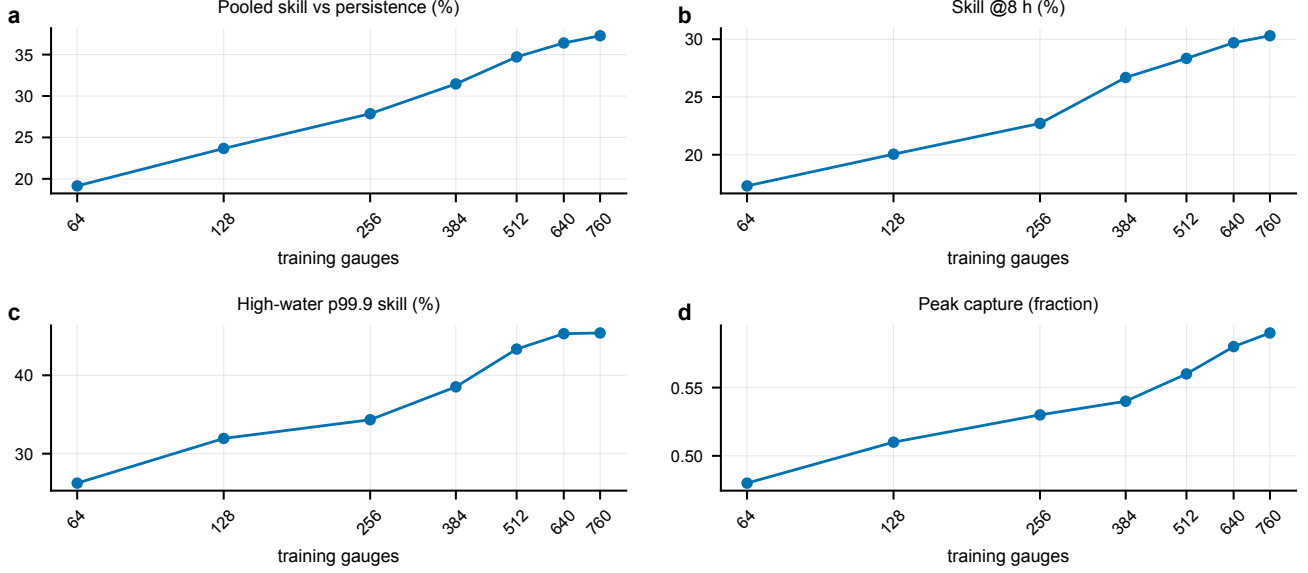


Figure 4: Data scaling for zero-shot storm surge forecasting on the fixed 84-gauge marine test set: seven training-set sizes (64 to 760 gauges, identical recipe; the largest is the delivered model). (a) Pooled skill against persistence, (b) skill at 8 h, (c) high-water p99.9 skill, (d) peak capture. Point and extreme-event skill rise across the tested range, with the smallest gains between the largest sizes; conditional tail coverage and envelope quality scale in parallel with small wobbles at the largest sizes of the size seen in the replicate run (Extended Data Table 3).

Table 2: What local information adds. (a) The local-information ladder: zero-shot NNSE at 8 h on the 84 marine test gauges as local information is added to the gauge-free configuration. (b) Head-to-head against the hydrodynamic baseline at equal local information: the ERA5-forced GTSM surge reanalysis (nearest output point, median 1.5 km; high-passed with the same 30-day filter as the target) on identical 48-h continuous windows over the 2010–2018 overlap (80 gauges; storm columns over the 61 gauges with at least five p99.9 windows). Tier A gives neither side a local observation; tier B gives both the observation at issue time (the reanalysis through a 25-h mean-error bias correction, the forecast through its persistence anchor). The two systems read different meteorological inputs (ERA5 fields for the reanalysis, the reanalysis-forcing series at the gauge for the neural network; Methods). RMSE in cm; wins = gauges at which the neural network has the lower 8-h RMSE; bold = the block-bootstrap interval of the paired difference within the tier excludes zero. NNSE at 8 h: gauge-free 0.60 and GTSM 0.62 (tier A); neural network 0.80, GTSM with bias correction 0.67, persistence 0.67 (tier B).

a

Configuration	local information added	NNSE@8h
All open products (EOT20 + reanalysis forcing)	none	0.593
+ local tidal statics	harmonic-fit constants	0.598
+ measured surge scale	one scalar σ_s	0.618
+ real-time water level (flagship)	208 h local history	0.782

b

		full distribution @8h	full distribution @48h	wins @8h	p99.9 storm windows @8h	p99.9 storm windows @48h	peak capture
A: no gauge	gauge-free model	6.7	6.7	38/80	11.8	13.3	0.57
	GTSM	6.2	6.1		8.8	9.9	0.61
B: gauge at t_0	neural network	4.2	6.0	66/80	7.9	11.2	0.62
	GTSM + bias correction	5.5	6.3		7.7	9.6	0.66
	persistence	5.9	10.1		12.4	23.8	0.34

network is better on the bulk at short leads and no difference is detectable inside storms.

The 99th-percentile envelope is well calibrated hour by hour but covers fewer than half of storm peaks

Point forecasts under-predict the sharpest peaks at long leads for a structural reason: conditional-mean regression against unpredictable peak timing. The quantile envelopes carry part of that information. At the true peak hour of p99.9 events, the mean q99 height is 0.91 to 0.99 of the mean peak height in the deployment setting at every peak-lead bucket including 25 to 48 h (where the point forecast reaches 0.56 and persistence 0.12); across the four rotations that ratio ranges from 0.69 (Europe) to 0.93 (Japan), so envelope quality, like point skill, degrades with regime distance. The ratio is a mean and hides the spread: the fraction of storm peaks that lie under the q99 envelope is 41% in deployment and 23 to 46% across the rotations, highest for peaks within 6 h of issue and lowest at 25 to 48 h. Overall q99 coverage is close to its nominal value of 0.99, the fraction of hours a 99th-percentile envelope should contain (0.98 in deployment, 0.96 to 0.98 across the rotations) while q90 under-covers (0.88 in deployment and 0.83 to 0.87 across the rotations, against 0.90). Conditional tail coverage, the share of high-water hours under the q99 envelope (a diagnostic, since even a calibrated marginal quantile sits below nominal on selected hours⁴⁰), is 0.75 in deployment and 0.52 to 0.74 across rotations, which marks headroom for conformal or tail-specific recalibration. Envelope crossing is zero throughout, and under the strictly proper pinball score the quantile trajectories beat per-gauge climatological quantiles by +45% at $\tau = 0.90$ (82 of 84 gauges) and +47% at $\tau = 0.99$ (77 of 84), so the envelopes carry conditional information beyond their width.

Discussion

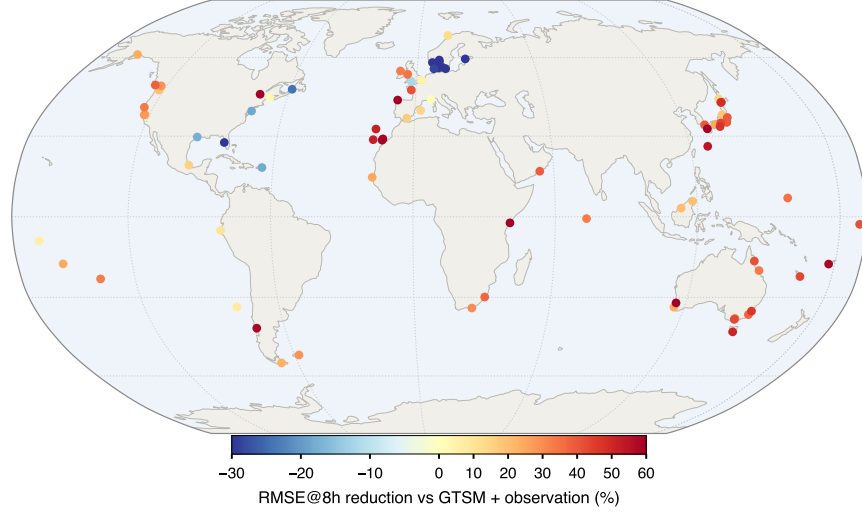
Three results carry the paper. First, one neural network trained on the global gauge network forecasts surge at gauges and whole regions that contributed no training data, and it keeps four fifths of its skill when driven by real weather forecasts. Second, transfer depends on which coastal regimes the training gauges cover as well as on how many there are: enlarging a Europe-free training set from 299 to 593 gauges lifted European pooled skill from zero to +11%, and at a fixed 298 gauges a 49-country training set transfers to Europe where a two-country training set does not. Third, the local water-level feed is the largest single term in the information ladder, the hydrodynamic model keeps the lower storm-window error where no gauge exists, and

the probabilistic envelope is calibrated hour by hour yet covers fewer than half of storm peaks.

Transfer is a property of the training network, on two axes. Gauge count raises accuracy, about five points of pooled skill per doubling, and the gains hold when the number of optimizer steps is held fixed (matched-update control, Methods). Regime diversity enables transfer, as the Europe-free intervention and the equal-count composition experiment show (Results). The two results together make regime coverage, alongside count, the property of the training network that transfer depends on. Further, we find that a gauge’s value is largest in the following cases: within the context of its own coast, where the real-time feed is the single largest term in the information ladder, and second, to each coast sharing its regime, where it enters as training diversity. The same result argues for finishing what GESLA began⁴¹: a single, quality-controlled global gauge archive; Extended Data Fig. 5 maps the record lengths of the present training network. Most of the effort in this study went into removing duplicate records, screening data quality, and masking tsunami hours, work that a consolidated archive would do once for everyone.

For a coastal service the results separate what the local observation contributes from what the network contributes. At a gauge with a water-level feed but no locally developed model, which is the case at many of the world’s several thousand operational gauges⁴¹, a model trained on the network provides, from open data and a single forward pass, hourly point forecasts that beat persistence by 37% in pooled error (30% with real forecast forcing) and by 32 to 40% inside storm windows, together with q99 envelopes whose hourly coverage is near nominal out to two days and which lie above fewer than half of the largest storm peaks. Compared with the global hydrodynamic reanalysis when both are given the same water-level observation, the model is more accurate over all hours at an eight-hour lead, and the two are statistically indistinguishable during storms. When neither has a local observation, the two are indistinguishable over all hours and the physics model is more accurate during storms, above all on wide shallow shelves. On the 80 gauges of this comparison the observation adds more skill at 8 h than the hydrodynamic model does: the gauge raises the neural network’s NNSE from 0.60 to 0.80 (tier A to tier B), while the corrected physics added to persistence leaves bulk NNSE at 0.67. Conditional tail coverage of 0.5 to 0.75 means the envelope misses between one high-water hour in four (deployment) and one in two (the weakest withheld region), so the outputs are candidates for evaluation as supplementary guidance alongside validated local systems, at the thresholds and lead times a service uses. Because the model is driven by weather forecasts, its skill in operation will rise and fall with the quality

Zero-shot forecast vs ERA5-forced GTSM reanalysis, information matched (80 gauges, identical windows, 2010-2018)
a bulk accuracy, both sides given the 10 observation: forecast better at 66/80 gauges



b storm peak magnitude (p99.9 windows): forecast better at 28/61 gauges

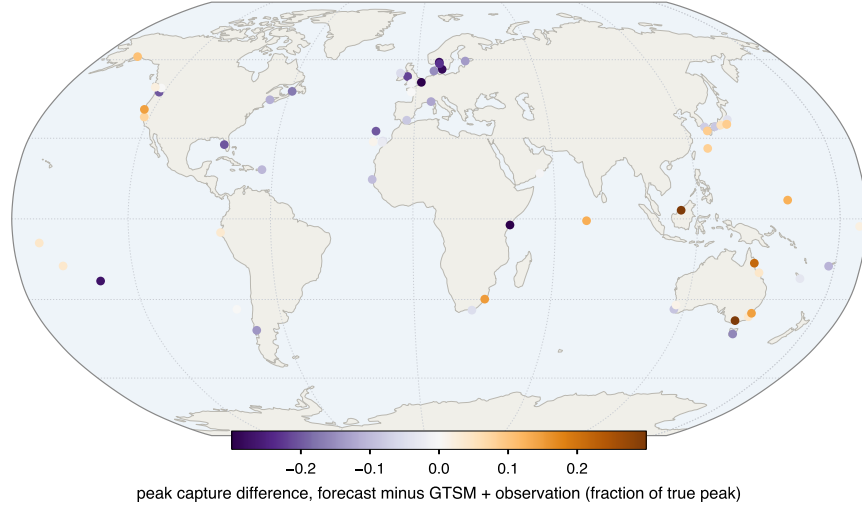


Figure 5: The head-to-head of Table 2b, mapped (tier B, same observation on both sides). (a) Per-gauge reduction of 8-h RMSE relative to the bias-corrected reanalysis: the neural network is better at 66 of 80 gauges; the exceptions concentrate in the Baltic and North Sea, where the physics has 22 to 46% lower error at 8 of 9 gauges. (b) Peak-capture difference inside p99.9 storm windows (61 gauges with at least five windows): the physics is better at 33 of 61 gauges, a difference whose interval spans zero (Table 2b).

of those forecasts.

It is important to highlight the differences between development of a forecast, operational or not, and the development of an early warning system^{42–44}. The former does not imply the latter, and without properly describing the policies, actions and decisions that may be influenced by such a system, it is not appropriate to retrofit statements of usefulness or value retrospectively based on the existence of a sufficiently skilful forecast (from a statistical perspective) or what is perceived to be an “operational” EWS. This is increasingly important as extreme event risk continues to evolve, leading to some areas being at risk that have not experienced

certain types of hazards in the past^{45,46}. An operational EWS for storm surge can yield benefit for some communities and sectors, yet still lead to zero benefit, and perhaps cause harm (for example, in terms of creating a distraction and leading to reprioritization away from the most vulnerable to those that can act when receiving such messaging)⁴⁷.

Therefore, even with the most advanced EWS, the areas that integrate the most data (of sufficient quality) are the ones best placed to benefit, and whether that benefit is realized depends on the dissemination, action capacity, governance and financing described above. These are generally the countries in higher socioeco-

505 nomic categories, so uneven data and technical capacity may widen the gap between the availability of an improved forecast and its accessibility and usefulness in the most vulnerable communities^{48,49}.

That said, the potential of “widening the gap” of 510 storm surge impacts between wealthier and lower income areas should not stop progress. Rather, all should be aware of this dynamic, and seek alignment with the socioeconomic factors in the areas such a system will be both present in and be available at a sufficiently 515 high level of quality or skill. These are two of the enabling environments of developing EWS of any type, with others including governance structures, existence of standard approaches to monitoring, evaluation and revision⁵⁰, and financing to allow for anticipatory actions to take place, for all, when the EWS indicates, 520 for example, a critical height of storm surge, to occur at certain locations inland, for a critical period of time. In the terms of this paper, shared observations supply the transferable forecast; whether it becomes a warning 525 depends on these local conditions.

The limits are as concrete as the gains. Most skill is measured under perfect-prognosis forcing; two full years of real 48-h forecast forcing keep four fifths of the pooled skill and beat persistence at 72 of 79 gauges, but 530 the largest peaks at 27 to 45 h of lead lose a quarter of their point capture (0.47 to 0.36), and the reforecast is an older, coarser system than current operational NWP, so an evaluation with a current ensemble remains to be done. Peak magnitude is under-predicted by the point 535 forecast, the q99 envelope lies under the true peak in more than half of the largest storms, and the physics model has the lower storm-window error where no gauge is available, above all on shallow shelves and in the semi-enclosed Baltic, where the information sits in the spatial 540 wind field; spatial forcing context and tail-specific heads are the identified levers. Surge is one component of coastal flooding; wave run-up and rain-driven flooding are outside this paper. Splits are spatial, so test-gauge years overlap training years at other gauges, and three 545 per-gauge scalars used at inference are computed over the gauge’s full record (Methods).

Global models built on simulation^{23,26,51} attack the same coverage gap from the other side, with dense reanalysis and hydrodynamic fields as inputs and either 550 observed gauges²³ or synthetic storm ensembles^{26,51} as targets, at the price of registered data products. The two approaches are complementary: observations give calibrated local truth where gauges exist and measured transfer estimates where they do not, and simulation 555 gives spatial density between gauges; a learned post-processor that ingests the reanalysis is the obvious hybrid and would reintroduce the registered product the rest of this pipeline avoids. On architecture, a 6M-parameter recurrent model with per-hour forcing 560 led a 23M-parameter attention model at 71 of 77 test

gauges, and a generic time-series foundation model with no forcing beat persistence only in the autocorrelation-dominated first hours (Methods; Extended Data Table 4); we read the result as a match between inductive 565 bias and the generating process, since surge integrates wind stress and pressure hour by hour, which is the computation a recurrent state performs, consistent with reports that generic foundation models lag domain models on streamflow⁵² and that simple models beat 570 transformers on long horizons⁵³.

The next steps follow from the map. Spatial forcing context, tested first in the Baltic, addresses the one regime where physics wins on the bulk; conformal recalibration on a small local sample or at region level 575 addresses tail coverage; and a prospective experiment in an unrepresented regime, adding a handful of gauges and re-scoring a withheld proxy gauge, would test the observing-system argument forward. We release the dataset recipe, splits, protocol, and models so that the 580 next claim in this space, ours or anyone’s, can be tested against a shared benchmark.

Methods

Sea-level observations and surge extraction

585 Water-level observations come from the GESLA-3 archive⁴¹. Native-resolution records are gross-clipped (10 m water level), flag-filtered (GESLA `use_flag=1`), and hourly averaged; surge values with $|s| \geq 4$ m are masked as non-meteorological or instrumental (hourly-mean 590 detided surges above 4 m at a gauge are, in this dataset, tsunamis, datum jumps, or sensor failures). Storm surge is extracted from the observed water level η by removing the harmonic tide (utide⁵⁴) and low-frequency variability,

$$\begin{aligned}\eta(t) &= T(t) + m(t) + s(t), \\ s(t) &= \eta(t) - \hat{T}(t) - \hat{m}(t),\end{aligned}\quad (1)$$

595 where $\hat{T}$ is a segmented harmonic fit (utide, 8-year blocks with Sa/Ssa and a linear trend per block), $\hat{m}$ is a 30-day moving mean, and the residual s is the surge signal after MAD-based despiking (8σ). Segmentation is essential for extremes: a single 101-year fit at 600 The Battery underestimates the peak surge of Sandy (2012) by 26% (2.59 m to 1.92 m), which the segmented fit removes. Detiding quality is audited twice. A coherent per-gauge harmonic re-fit gates admission (median coherent M2 residual 0.11 cm; gauges above 2 605 cm excluded). Because a coherent fit under-measures incoherent tidal-band energy, an independent spectral audit recomputed total M2-band energy on the processed surge: the fold-wide median is 1.0 cm, and every training-pool gauge above 5 cm is an estuary, river,

weir, or lake site (relabelled out of domain); three marine test gauges retain elevated residuals (Holyhead, the São Marcos Bay gauge in Brazil listed as Madeira, Puerto Montt, 5 to 12 cm) and are kept in the test set and flagged in the released tables. The same audit verified that landmark storm peaks survive cleaning where raw data exist (Sandy at The Battery, 2.59 m; Xynthia at Les Sables, 1.12 m; Haiyan at Cebu), that the segmented fit leaves only sub-p99 single-hour steps at block joints (median 4.8 cm at 0.0015% of hours), and that the 30-day high-pass attenuates 1-to-3-day surge pulses by only 4 to 11%. The high-pass deletes isolated data islands shorter than about 15 days and pre-hourly-sampling eras, so the 1953 North Sea flood survives in this dataset only at Esbjerg; the $|s| \geq 4$ m mask removes 128 fold hours, a few of which at German estuary gauges are real storm peaks (those gauges are out of domain). The low-frequency baseline and despiking windows are centred, so the surge targets are retrospectively defined, and the harmonic constants are full-record estimates, as is standard for detided-surge datasets. Recomputing all 84 marine test gauges with strictly trailing, causal windows and re-evaluating the delivered model zero-shot changes no conclusion (all evaluation windows): pooled skill moves from +39 to +36% (persistence itself strengthens, because slow anomalies remain in the causal target), NNSE at 8 h rises from 0.786 to 0.818, storm-window skill at 8 h from +35 to +40%, and 80 of 84 gauges still beat persistence. The centred definition is conservative on extremes: a centred 30-day window absorbs part of each storm peak into its baseline.

Forcing

Hourly 10 m wind (u , v), mean sea-level pressure, and precipitation at each gauge location from the OpenMeteo historical archive (open, no registration), requested with the API’s default product, which combines ERA5⁵⁵ at 0.25°, ERA5-Land at 0.1° over land, and ECMWF IFS analyses at 9 km from 2017 onward; we refer to it as reanalysis forcing. This is a perfect-prognosis setting: reanalysis stands in for a forecast-time atmospheric field, isolating the surge-mapping problem from NWP error, consistent with prior surge ML^{20,24}. Forcing fetches begin in 2000, so the median gauge contributes 44% of its surge record to the joined dataset (quartiles 27 to 73%, minimum 1%). For 251 of the 1,047 gauges with at least 3,000 joined hours the forcing fetch returned nulls for whole calendar years, so 23 to 78% of their post-2000 hours (median 54%) are dropped at the join; the remaining gauges lose at most 20%.

Cleaning beyond the standard pipeline

Two failure modes discovered during this work motivate cleaning steps we have not seen in prior surge ML. *Cross-provider duplicates.* The documented catalogue cleaning removes 74 lake and river stations, 23 gross-range failures, and 321 cross-provider duplicates (377 sub-2-km pairs, 42% of raw records); a final 0.02°-box dedupe in the split builder catches 26 residual pairs (for example Abashiri via UHSLC and via JMA) and keeps one member of each. Undetected duplicates straddling a train/test split would constitute direct leakage. *Seismic contamination.* Percentile-selected extreme events include tsunamis and seiches: the largest “surge” in our Japan test set (112 cm, Aburatsubo, March 1952) is the Tokachi-oki M8.1 tsunami week. We mask hours within graded windows of USGS M7+ events ($M \geq 7.0$ within 1,000 km: 72 h; $M \geq 7.5$ within 3,000 km: 96 h; $M \geq 8.0$ within 8,000 km: 168 h). The mask removes a median of 0.8% of hours (2.0 to 2.4% in Japan, 0.1% in Germany) but a far larger share of p99.9 exceedance hours where it matters: up to 45% at individual gauges (Vung Tau, Vietnam), with 16 fold gauges above 10% (Supplementary Table 3b). The mask entered the pipeline mid-study; all reported results are computed on masked records, and the rotations were retrained under the final validation-fold protocol after masking.

Dataset, splits, and pre-registration

The final dataset holds 1,054 gauges with joined observations and forcing. Removing the 100 pre-registered test gauges (84 marine primary, 16 out of domain), the 127 gauges inside their 0.5° exclusion buffers, one member of each of 23 remaining co-located pairs, and four gauges that the catalogue cleaning had already excluded (three lake, lagoon, or river sites and one co-located twin) leaves the 800-gauge split pool (760 training, 40 validation). Fold sizes are nominal: seven of the 1,054 gauges have fewer than 3,000 joined hours after masking, because their records end at or before the start of the forcing archive in 2000 or, in one case, lost their wind forcing to a download failure, and they contribute no windows; four sit in the training fold (756 effective training gauges), one in the validation fold (39 effective), one among the sixteen out-of-domain test gauges (15 effective), and one outside the split. The training-set sizes of the scaling and rotation experiments are nominal in the same way (effective sizes 63, 126, 254, 380, 508, 636, and 756 along the ladder; 592, 587, 606, and 785 for the Japan, Europe, North America, and Oceania rotations). Two evaluation settings were frozen before the corresponding training (split files in the repository). *Deployment (unseen gauge within the observed network).* 100 gauges were pre-registered, stratified across all continents, each with a 0.5° exclusion buffer (remaining gauges form

the training pool). Two audits then screened the pre-registered set by water-body type before the final models were evaluated: a composition audit relabelled six gauges (five Great-Lakes, one Antarctic ice-shelf station), and an independent detide-and-cleaning audit relabelled ten more whose records are estuarine rivers, a tidal weir, a delta channel, or a lake (Elbe, Loire, and Ems estuary chains, the Sacramento delta, the Columbia River, Lake Huron; Supplementary Fig. 1). The primary test set is therefore 84 marine open-coast gauges (Asia 18, Europe 18, North America 15, Oceania 10, Africa 8, South America 8, Pacific islands 7; grouping as in Table 1). The relabeling is conservative on the headline metric: on the dress-rehearsal model it raises pooled skill by 0.7 but lowers 8-h skill by 1.3 points, and that model beat persistence on all sixteen relabelled gauges; the delivered model beats persistence on 14 of the 15 that pass the joined-hours minimum (all windows; Supplementary Information). Pre-registration integrity, stated once and completely: the test set was evaluated twice, in a dress rehearsal with an earlier training set and in the final clean-protocol run; between the two, no architecture, hyperparameter, or selection decision changed in response to test results, and the protocol changes (validation-fold selection and the QC-clean training universe) were audit-driven and test-blind. The only change to the test set itself was the water-body relabeling, whose criterion is physical and whose effect is quantified in both directions above. One registered marine gauge, Varennes on the fluvial St Lawrence just downstream of Montreal, passed both audits (its tidal range is 3 cm, so the detide screen could not flag it) and stays in the primary set as registered; it is one of the three gauges that do not beat persistence (Supplementary Fig. 1a). The split files are shipped unchanged in the public repository; the commit history of the working repository from which the release is built records when each was frozen. The training pool retains 62 lake, river, and inner-estuary gauges (8% of the pool; Supplementary Fig. 1): heterogeneous forcing-response regimes are useful training signal even where they would contaminate evaluation. Two ablations confirm their effect on marine test skill is nil: removing the 54 of them in the Great Lakes, on the St Lawrence, and on the Dutch inland waterway from the deliverable (760 to 706 training gauges) moves pooled skill by 0.6 and 8-h skill by 0.3 points, and at matched count (640 gauges with proportional lake content against 640 marine-only) the differences are 0.5 and 0.3 points with mixed signs, all of the size seen in the replicate run. The field norm for this setting is no buffer at all^{25,34}; we additionally deduplicate at 2 km. *Out-of-network (unseen region)*. Four rotating leave-region-out splits (Japan and Pacific, 77 test gauges; Europe, 234 (233 evaluated, 231 with enough continuous windows); North America, 220 (214 evaluated); Oceania,

111), each evaluated by a model trained without any data from the region and selected on a 40-gauge validation fold carved from the remaining pool. Regions are defined by membership, so the minimum train-to-test separation differs by rotation (2.2° North America, 2.3° Japan, 6.4° Oceania, 7.5° Europe); no training gauge lies within 0.5° of a test gauge in any split. Gauges with fewer than 3,000 joined observation-forcing hours are excluded from evaluation. Rotation test sets contain every gauge of the region that passes the joined-hours minimum, including lake, river, and inner-estuary gauges that the marine deployment set excludes: 80 of the 214 evaluated North American gauges (the Great Lakes and their connecting channels, the St Lawrence above the head of tide, the Columbia River above its estuary, and the Sacramento–San Joaquin Delta), 7 of 231 European (the Elbe above Cuxhaven, the Loire above Saint-Nazaire, and the Rhine–Meuse tidal rivers), and 1 of 111 Oceanian (the Gippsland Lakes); the classification is released with the code. Restricted to marine gauges, the North American rotation scores +16% pooled and +14% at 8 h (NNSE 0.713; 110 of 134 gauges beat persistence; the 80 non-marine gauges score −4% pooled), Europe and Oceania change by 0.2 points or less, and the union becomes 414 of 545. The rotation splits predate the final 0.02° dedupe pass, so 16 cross-provider pairs within 2 km remain inside the test sets (Europe 7, North America 2, Oceania 7); dropping the higher-gap member of each pair changes pooled skill by at most 0.2 points and the union count to 465 of 617.

Model and training

An encoder LSTM (2×512) reads 208 h of context (surge, wind u and v , pressure, precipitation, with five static coastal attributes concatenated per step: sin/cos-encoded position, tidal range, tidal form factor). A decoder LSTM (2×512) rolls 48 hourly steps, each consuming that hour’s future forcing plus statics, initialized from the encoder state. A zero-initialized linear head outputs three residual trajectories relative to the last observation (the persistence anchor). For issue time t_0 and lead $\ell \in \{1, \dots, 48\}$,

$$\hat{y}^{(k)}(t_0 + \ell) = y(t_0) + \hat{\Delta}^{(k)}(\ell),$$

$$k \in \{\text{pt}, q90, q99\}, \quad (2)$$

so an untrained model reproduces persistence exactly and gradient descent learns only corrections to it. The point head is trained with magnitude-weighted MSE,

$$\mathcal{L}_{\text{pt}} = \frac{\sum_{\ell=1}^L w(y_{t_0+\ell}) (\hat{y}_{t_0+\ell}^{(\text{pt})} - y_{t_0+\ell})^2}{\sum_{\ell=1}^L w(y_{t_0+\ell})},$$

$$w(y) = 1 + |y|, \quad (3)$$

with y in units of per-gauge surge standard deviation, and each quantile head with the standard pinball loss

at level $\tau \in \{0.90, 0.99\}$,

$$\mathcal{L}_\tau = \frac{1}{L} \sum_{\ell=1}^L \max(\tau u_\ell, (\tau - 1) u_\ell),$$

$$u_\ell = y_{t_0+\ell} - \hat{y}_{t_0+\ell}^{(\tau)}, \quad (4)$$

combined as $\mathcal{L} = \mathcal{L}_{\text{pt}} + \lambda(\mathcal{L}_{0.90} + \mathcal{L}_{0.99})$ with $\lambda = 1$. The model has 6.35M parameters. Normalization: per-window standardization of context surge, per-gauge standardization of forcing, targets in units of per-gauge surge standard deviation. Training: AdamW (weight decay 0.05), learning rate 10^{-3} with cosine decay and a 10% linear warmup, 12 epochs, batch 64, gradient clip 1.0; checkpoint selection by zero-shot point RMSE at 8 h on a 40-gauge validation fold carved from the training pool (stratified, excluded from training), never on the test fold and never by training loss; for the delivered model the selected checkpoint is the final epoch. The persistence-anchored residual repairs an otherwise universal mean-regression failure; statics enable cross-site generalization⁵⁶; per-hour forcing consumption in the decoder is the decisive architectural factor (below); the quantile head adds calibrated envelopes and improves the point forecast on extremes through shared-trunk multi-task learning. The gauge-free variant is the same architecture retrained with the water-level context channel and the anchor set to zero; its tidal statics come from EOT20³⁹ and its output scale from a gradient-boostered regression of σ_s on reanalysis-forcing covariates fitted on training gauges (log-correlation 0.85, median error 14%; EOT20 reproduces local tidal range with log-correlation 0.93 and median error 3%).

Evaluation protocol and metrics

Every reported number comes from a single evaluation script run on the complete test sets (Supplementary Table 1 maps every reported quantity to its implementation). The persistence reference at issue time t_0 is $\hat{y}_{\text{per}}(t_0 + \ell) = y(t_0)$ for all ℓ . Per-gauge, per-lead RMSE and the corresponding skill are

$$\text{RMSE}_s(\ell) = \left[\frac{1}{|\mathcal{T}_s|} \sum_{t_0 \in \mathcal{T}_s} (\hat{y}_s(t_0 + \ell) - y_s(t_0 + \ell))^2 \right]^{1/2},$$

$$\text{skill}(\ell) = 1 - \frac{\langle \text{RMSE}_s(\ell) \rangle_s}{\langle \text{RMSE}_s^{\text{per}}(\ell) \rangle_s}, \quad (5)$$

with $\langle \cdot \rangle_s$ the mean over test gauges. Pooled RMSE aggregates squared errors over all leads $\ell \leq L$ before the root,

$$\text{RMSE}_s^{\text{pool}} = \left[\frac{1}{L} \sum_{\ell=1}^L \text{RMSE}_s(\ell)^2 \right]^{1/2}, \quad (6)$$

which makes its failure mode explicit: because $\text{RMSE}_s^{\text{per}}(\ell)$ grows roughly monotonically with ℓ (surge decorrelation), the sum is dominated by the longest

leads, and a model can beat persistence pooled while losing to it at every operational lead $\ell \leq 12$. An early attention-based variant of ours did exactly this: +14% pooled skill while losing to persistence at 4 h (−47%), 8 h (−15%), and 12 h (−8%), capturing 16% of p99.9 peak height and under-predicting 100% of peaks. Since aggregate metrics of this type are the field’s default, we report per-lead and event-conditioned metrics throughout, extending to surge the extreme-evaluation critique established for global weather models^{57,58}. NNSE is $1/(2 - \text{NSE})$ with NSE the Nash–Sutcliffe efficiency, computed per gauge at a fixed lead and averaged. Event-conditioned metrics use station-level thresholds (p99/p99.5/p99.9 of the surge distribution over the gauge’s full test record): event-window and high-water-hour RMSE (the latter pooled over hours across gauges), per-lead skill within event windows, and peak capture at the true peak hour t^* of each p99.9 window,

$$\text{capture} = \frac{\langle \hat{y}(t^*) \rangle_{\text{windows}}}{\langle y(t^*) \rangle_{\text{windows}}}, \quad (7)$$

reported overall and stratified by the lead position of the peak. Quantile heads are scored by coverage, overall and conditional on high water,

$$\text{cov}_\tau = \Pr(y \leq \hat{y}^{(\tau)}),$$

$$\text{cov}_\tau^{\text{tail}} = \Pr(y \leq \hat{y}^{(\tau)} \mid y \geq p99_s), \quad (8)$$

by sharpness (mean envelope width $\hat{y}^{(q99)} - \hat{y}^{(\text{pt})}$ on storm against calm hours, reported as their ratio), by the quantile-crossing rate $\Pr(\hat{y}^{(q90)} > \hat{y}^{(q99)})$, and by the pinball score against per-gauge climatological quantiles. Because neighbouring gauges are not independent, uncertainty is reported with a spatial-block bootstrap: gauges are grouped into eleven ocean regions (and, as a check, by country), regions are resampled with replacement 4,000 times, and each skill or paired difference is recomputed on every resample; the 2.5 and 97.5 percentiles are quoted as 95% intervals. Deployment pooled skill is +37% with interval +33 to +41% and 8-h skill +30% (+26 to +34%); the withheld-region experiments yield +30% (+18 to +49%) for Japan, +11% (0 to +22%) for Europe, +10% (+1 to +28%) for North America, and +8% (+8 to +25%) for Oceania. These intervals are wide because a withheld-region experiment contains only three to four ocean regions, so the block bootstrap has few blocks to resample; the width measures the resampling design and carries no information about the model. Paired comparisons are reported as counts of gauges together with the interval of the mean difference; bold entries in the tables mark intervals that exclude zero.

Baselines

Persistence. The persistence baseline holds the last observed surge; because the model’s head is zero-initialized and its output is anchored on the same observation, persistence is the model’s own zero-correction output, so both sides always see the same t_0 sample.

Hydrodynamic reanalysis. The ERA5-forced GTSM v3 surge reanalysis (10-min, hourly-averaged) was sampled at the nearest output point to each gauge (median 1.5 km; three gauges with the nearest point beyond 25 km change no conclusion), high-passed with the same centred 720-h mean as the target, and mean-aligned per gauge; 80 gauges have sufficient joint coverage over 2010–2018, 61 of them with at least five p99.9 storm windows among the continuous windows. In tier B it receives the observation through a bias correction equal to its mean error over the 25 h preceding issue time (the 25-h mean removes a semidiurnal component present in the gauge-minus-reanalysis difference); instantaneous anchoring and a mean correction over the full 208-h context are both weaker (6.3 against 5.5 cm at 8 h for the 208-h correction) and are reported in the repository. Without the high-pass, 22% of the reanalysis’s apparent 8-h error is low-frequency signal that the target definition removes. For context, uncalibrated GTSM at gauge locations scored NNSE 0.195 in the simulation-based study²³, and CoDEC-ERA5 validates against GESLA annual maxima at 485 gauges with an RMSE of 0.26 m¹⁹. Persistence, the gauge-free model, and the flagship are scored on the identical windows. Storm windows in this comparison are defined by the p99.9 of the window targets over 2010–2018, the same threshold for both systems. The comparison matches the local observation given to each side and nothing else: the reanalysis is driven by ERA5 wind and pressure fields over its whole domain, the neural network by the Open-Meteo forcing series at the gauge, so the head-to-head compares two complete forecasting systems that differ in meteorological input as well as in method.

Simulation-based densification. The comparison to the published NNSE 0.556²³ is indirect: their metric is pooled across gauges, their test targets are single hours drawn preferentially at cyclone landfall times, their target is detided differently, and their test split is spatial and temporal; ours is per gauge, all hours, spatial only.

Generic time-series foundation model. Chronos-Bolt⁵⁹, a general-purpose time-series forecasting model pretrained on large collections of unrelated time series (we use its small variant), forecasts from the 208-h surge history alone, with no weather input and no training

on our data, on the same windows and gauges as the architecture comparison.

Architecture factor study (Extended Data Table 4). On a 298-gauge training set with Japan withheld (77 test gauges), all rows under the clean protocol: the recurrent model; an attention model (23M parameters) whose predictor cross-attends from 48 hourly queries to the encoded context and to 192 hourly forcing tokens that carry learned hour and channel embeddings, trained with the same loss and budget at the transformer-conventional learning rate of 3×10^{-4} (no learning-rate sweep on either side); and Chronos-bolt. The recurrent model has the lower 8-h RMSE at 71 of 77 gauges.

Real-forecast forcing

GEFSv12 reforecast control-member 10 m winds, mean sea-level pressure, and accumulated precipitation (0.25° , 3-hourly, issued at 00 UTC only) were fetched by HTTP byte range from the public archive and sampled at the nearest grid point to each gauge (scripts in the repository). For the season-scale check every 00 UTC cycle of 2012 and 2017 was used (731 cycles; the two calendar years within the reforecast period with the most test-gauge coverage, 72 and 65 of the 84 gauges with at least 80% of hours, chosen before any forecast was scored). At each cycle the model reads the same 208 h of observed context and reanalysis forcing as in deployment and receives, for the 48-h horizon, either the reanalysis forcing (perfect prognosis) or the reforecast: 3-hourly values interpolated to hourly and anchored on the reanalysis value at issue time, with the 3-h precipitation buckets rebuilt from the reforecast’s alternating 3-h and 6-h accumulations, then normalized with the gauge’s reanalysis statistics without bias correction. Windows require a fully continuous 256-h record; persistence, the reanalysis-forced and the reforecast-forced runs are scored on identical windows with the metrics of the main evaluation (per-lead RMSE pooled over hours for the skill curves, gauge-mean RMSE for the pooled and fixed-lead skills, storm windows defined by the gauge’s full-record p99.9 threshold). For the storm-case check, twelve of the largest test-gauge storms were selected before any metric was seen and nine retain complete forecast and observation coverage (eight storms, Xavier at two gauges); the issue time is the 00 UTC cycle 24 to 48 h before the catalogued peak, so that the peak falls in the second day of the horizon (implemented as the cycle 36 to 60 h before the peak, moved one day later for the three storms where that cycle leaves the peak outside the horizon), which gives true peak leads of 27 to 45 h (Supplementary Table 2). Context and persistence are identical between the two forcings, so the difference isolates the forcing.

Data integrity notes

A single same-pipeline replicate (512 training gauges, unseeded, clean protocol) differs from its twin by 0.8 points pooled skill, 0.8 points at 8 h, 0.004 NNSE, and 0.01 peak capture (0.6, 0.2, 0.002, and 0.02 on all windows); one replicate does not bound run-to-run variation, and differences of this order should not be interpreted. A matched-update control retrained the 64- and 256-gauge training sets to at least the 760-gauge run's 77,028 optimizer steps (77,077 and 77,922 steps in whole epochs): the 64-gauge model moved from +19 to +20% pooled and from +17 to +19% at 8 h (validation-selected checkpoint at epoch 18 of 143), the 256-gauge model from +28 to +27% pooled with 8-h skill unchanged at +23%. There is no temporal test split: three per-gauge scalars used at inference (the surge scale σ_s , the forcing mean and standard deviation, and the tidal statics, the last fitted on the gauge's most recent ten years) are computed over the gauge's full record, evaluation hours included; the information ladder measures what removing them costs. Each evaluation case is a window of 256 consecutive rows of the record, 208 hours of context followed by 48 hours to forecast, taken every 48 hours along the record. Because hours with missing data are removed before the windows are taken, 28.7% of the 239,757 deployment windows contain a gap somewhere in their 256 rows (9.0% inside the 48-h forecast period, 2.2% at the 8-h lead). Across a gap a persistence forecast spans more real time than its nominal lead, which favours the model on those windows. Training windows are taken the same way. All primary results are therefore computed on the fully continuous windows (71% of deployment windows, 55 to 80% across the rotations; gauges with fewer than 20 such windows are dropped, which removes two European rotation gauges), and Extended Data Table 5 reports every headline number on both sets: on all windows pooled skill is 2 points higher in deployment and 2 to 5 points higher in the rotations, 8-h skill 0 to 3 points higher, no comparison changes sign, and one deployment gauge (Veracruz) moves from +0.5 to -18% pooled skill, so the count of gauges beating persistence is 82 of 84 on all windows and 81 of 84 on continuous windows. Two known defects remain in the training data: one training gauge (Dublin Port, Ireland) retains part of a segment in which the sensor reported a constant value, and one training gauge lost its wind forcing to a download failure and contributes no windows. Neither gauge is in the test set. The seismic mask is rule-based; Supplementary Table 3 and Supplementary Fig. 3 document its behaviour on four events (Tohoku 2011 and Sumatra 2004, removed; Sandy 2012 and Haiyan 2013, retained) and list the gauges whose extreme hours it affects most.

Data and code availability

All code, split files, per-gauge evaluation tables, the training logs of the seven runs made after run logging was added, the sampled reforecast forcing at the test gauges, and the cleaned-dataset recipe: <https://github.com/HeshanChen/surge-zeroshot>; the release of record (v1.0.0) is archived at Zenodo (<https://doi.org/10.5281/zenodo.22787673>). GESLA-3 and the Open-Meteo archive are open; no registered products are required anywhere in the forecasting pipeline. The hydrodynamic baseline is distributed via the Copernicus Climate Data Store and was used for comparison only.

Acknowledgements and author contributions

H.C. designed and carried out the study and wrote the manuscript; A.K. framed the work within humanitarian practice and revised the manuscript. This work received no dedicated funding. The authors declare no competing interests.

References

- [1] Edward N. Rappaport. Fatalities in the United States from Atlantic tropical cyclones: New data and interpretation. *Bulletin of the American Meteorological Society*, 95(3):341–346, 2014. doi: 10.1175/BAMS-D-12-00074.1.
- [2] Jochen Hinkel, Daniel Lincke, Athanasios T. Vafeidis, Mahe Perrette, Robert J. Nicholls, Richard S. J. Tol, Ben Marzeion, Xavier Fettweis, Cezar Ionescu, and Anders Levermann. Coastal flood damage and adaptation costs under 21st century sea-level rise. *Proceedings of the National Academy of Sciences*, 111(9):3292–3297, 2014. doi: 10.1073/pnas.1222469111.
- [3] Ebru Kirezci, Ian R. Young, Roshanka Ranasinghe, Sanne Muis, Robert J. Nicholls, Daniel Lincke, and Jochen Hinkel. Projections of global-scale extreme sea levels and resulting episodic coastal flooding over the 21st century. *Scientific Reports*, 10:11629, 2020. doi: 10.1038/s41598-020-67736-6.
- [4] World Meteorological Organization and United Nations Office for Disaster Risk Reduction. Global status of multi-hazard early warning systems 2024. Technical report, WMO, Geneva, 2024. <https://library.wmo.int/records/item/69085-global-status-of-multi-hazard-early-warning-systems-2024>.

- [5] Alexandre K. Magnan, Michael Oppenheimer, Matthias Garschagen, Maya K. Buchanan, Virginie K. E. Duvat, Donald L. Forbes, James D. Ford, Erwin Lambert, Jan Petzold, Fabrice G. Renaud, Zita Sebesvari, Roderik S. W. van de Wal, Jochen Hinkel, and Hans-Otto Pörtner. Sea level rise risks and societal adaptation benefits in low-lying coastal areas. *Scientific Reports*, 12:10677, 2022. doi: 10.1038/s41598-022-14303-w.
- [6] Andrew Kruczkiewicz, Fabio Cian, Irene Monasterolo, Giuliano Di Baldassarre, Astrid Caldas, Moriah Royz, Margaret Glasscoe, Nicola Ranger, and Maarten van Aalst. Multiform flood risk in a rapidly changing world: what we do not do, what we should and why it matters. *Environmental Research Letters*, 17(8):081001, 2022. doi: 10.1088/1748-9326/ac7ed9.
- [7] Jonathan Flowerdew, Kevin Horsburgh, Chris Wilson, and Ken Mylne. Development and evaluation of an ensemble forecasting system for coastal storm surges. *Quarterly Journal of the Royal Meteorological Society*, 136(651):1444–1456, 2010. doi: 10.1002/qj.648.
- [8] Ian L. Turner, Christopher K. Leaman, Mitchell D. Harley, Mandi C. Thran, Daniel R. David, Kristen D. Splinter, Nashwan Matheen, Jeff E. Hansen, Michael V. W. Cuttler, Diana J. M. Greenslade, Stefan Zieger, and Ryan J. Lowe. A framework for national-scale coastal storm hazards early warning. *Coastal Engineering*, 192:104571, 2024. doi: 10.1016/j.coastaleng.2024.104571.
- [9] T. Fernández-Montblanc, M. I. Voutsoukas, P. Ciavola, E. Voukouvalas, L. Mentaschi, G. Breyiannis, L. Feyen, and P. Salamon. Towards robust pan-European storm surge forecasting. *Ocean Modelling*, 133:129–144, 2019. doi: 10.1016/j.ocemod.2018.12.001.
- [10] Diane P. Horn. Storm surge warning, mitigation, and adaptation. In Jean T. Ellis and Douglas J. Sherman, editors, *Coastal and Marine Hazards, Risks, and Disasters*, pages 153–180. Elsevier, Amsterdam, 2015. doi: 10.1016/B978-0-12-396483-0.00006-6.
- [11] Jasper Verschuur, Nicola Ranger, and Jim W. Hall. Climate adaptation finance: From paper commitments to climate risk reduction. *Science*, 388(6754):1374–1377, 2025. doi: 10.1126/science.adx1950.
- [12] Siqi Zheng, Alex Jingwei He, and Yujie He. Authority versus competition: Intergovernmental interactions and subnational policy adoption in climate policy. *Review of Policy Research*, 43(3):e70056, 2026. doi: 10.1111/ropr.70056.
- [13] Elisabeth Stephens, Erin Coughlan de Perez, Andrew Kruczkiewicz, Emily Boyd, and Pablo Suarez. Forecast-based action. Technical report, Red Cross Red Crescent Climate Centre and University of Reading, 2015.
- [14] Ilan Kelman and Carina J. Fearnley. From multi-hazard early warning systems (MHEWS) to all-vulnerability warning systems (AVWS). *iScience*, 28(7):112977, 2025. doi: 10.1016/j.isci.2025.112977.
- [15] Renan Braga Ribeiro, Alexandra Franciscatto Penteado Sampaio, Matheus Souza Ruiz, José Chamber Leitão, and Paulo Chamber Leitão. First approach of a storm surge early warning system for Santos region. In *Climate Change in Santos Brazil: Projections, Impacts and Adaptation Options*, pages 135–157. Springer, Cham, 2018. doi: 10.1007/978-3-319-96535-2-7.
- [16] Henrike Brecht, Susmita Dasgupta, Benoit Laplante, Siobhan Murray, and David Wheeler. Sea-level rise and storm surges: High stakes for a small number of developing countries. *The Journal of Environment & Development*, 21(1):120–138, 2012. doi: 10.1177/1070496511433601.
- [17] Jiayi Fang, Shao Sun, Peijun Shi, and Jing'ai Wang. Assessment and mapping of potential storm surge impacts on global population and economy. *International Journal of Disaster Risk Science*, 5(4):323–331, 2014. doi: 10.1007/s13753-014-0035-0.
- [18] Lara J. Cushing, Yang Ju, Seigi Karasaki, Scott Kulp, Nicholas Depsky, Alique Berberian, Jessie Jaeger, Benjamin Strauss, and Rachel Morello-Frosch. Sea level rise and flooding of hazardous sites in marginalized communities across the United States. *Nature Communications*, 16: 9711, 2025. doi: 10.1038/s41467-025-65168-2.
- [19] Sanne Muis, Maialen Irazoqui Apecechea, Job Dul-laart, Joao de Lima Rego, Kristine Skovgaard Madsen, Jian Su, Kun Yan, and Martin Verlaan. A high-resolution global dataset of extreme sea levels, tides, and storm surges, including future projections. *Frontiers in Marine Science*, 7:263, 2020. doi: 10.3389/fmars.2020.00263.
- [20] Timothy Tiggeoven, Anais Couasnon, Chiem van Straaten, Sanne Muis, and Philip J. Ward. Exploring deep learning capabilities for surge predictions in coastal areas. *Scientific Reports*, 11:17224, 2021. doi: 10.1038/s41598-021-96674-0.
- [21] Tim H. J. Hermans, Chiheb Ben Hammouda, Simon Treu, Timothy Tiggeoven, Anais Couasnon,

- Julius J. M. Busecke, and Roderik S. W. van de Wal. Computing extreme storm surges in europe using neural networks. *Natural Hazards and Earth System Sciences*, 25:4593–4612, 2025. doi: 10.5194/nhess-25-4593-2025.
- [22] Michael Tadesse, Thomas Wahl, and Alejandra Cid. Data-driven modeling of global storm surges. *Frontiers in Marine Science*, 7:260, 2020. doi: 10.3389/fmars.2020.00260.
- [23] Patrick Ebel, Brandon Victor, Peter Naylor, Gabriele Meoni, Federico Serva, and Rochelle Schneider. Implicit assimilation of sparse in situ data for dense & global storm surge forecasting. In *CVPR Workshops (EarthVision)*, pages 471–480, 2024. arXiv:2404.05758.
- [24] M. Nagaraj, A. Rodriguez, and T. Wahl. Regional modeling of storm surges using localized features and transfer learning. *Journal of Geophysical Research: Machine Learning and Computation*, 2(3): e2025JH000650, 2025. doi: 10.1029/2025JH000650.
- [25] Samuel Daramola, David F. Munoz, Paul Munoz, Siddharth Saksena, and Jennifer L. Irish. Predicting the evolution of extreme water levels with long short-term memory station-based approximated models and transfer learning techniques. *Water Resources Research*, 61(3):e2024WR039054, 2025. doi: 10.1029/2024WR039054.
- [26] Benjamin Pachev, Prateek Arora, Jinpai Zhao, and Eirik Valseth. Global location-invariant peak storm surge prediction, 2026. arXiv:2603.25978.
- [27] Zesheng Liu, Doyup Kwon, Ning Lin, and Maryam Rahnemounfar. PACT: Peak-aware cross-attention graph transformers for efficient storm-surge emulation, 2026. arXiv:2605.09036.
- [28] Nils Melsom Kristensen, Mateusz Matuszak, Paulina Tedesco, Ina Kristine Berentsen Kullmann, and Johannes Röhrs. Flo: A data-driven limited-area storm surge model, 2026. arXiv:2601.02090.
- [29] Noujoud Nader, Stefanos Giaremis, Clint Dawson, Carola Kaiser, Karame Mohammadiporshokoo, and Hartmut Kaiser. StormNet: Improving storm surge predictions with a GNN-based spatio-temporal offset forecasting model, 2026. arXiv:2604.20688.
- [30] Lianjun Yang, Taoyong Jin, and Weiping Jiang. ASM-SS: the first quasi-global high-spatial-resolution coastal storm surge dataset reconstructed from tide gauge records. *Earth System Science Data*, 17(6):2793–2807, 2025. doi: 10.5194/essd-17-2793-2025.
- [31] Xiaohui Li, Xinhai Han, Jingsong Yang, Guoqi Han, Jiuke Wang, Lotfi Aouf, Shaoliang Peng, and Dake Chen. From regional to global: a typhoon-aware adaptive graph transformer for cross-basin tropical cyclone-induced storm surge forecasting, 2026. Preprint, Research Square, doi:10.21203/rs.3.rs-10797331/v1.
- [32] Angel Amores, Roberto Alcaraz, Marta Marcos, Philip Thompson, Ariadna Martín, and Thomas Wahl. MAROR: a multipoint autoregressive network for forcing-driven, uncertainty-calibrated reconstruction of hourly coastal sea level from 1940 to the present, 2026. Preprint, EGU sphere, doi:10.5194/egusphere-2026-5073.
- [33] NOAA/NWS Meteorological Development Laboratory. P-Surge: Probabilistic hurricane storm surge. Operational product, <https://slosh.nws.noaa.gov/psurge>, 2008.
- [34] Frederik Kratzert, Daniel Klotz, Mathew Herrnegger, Alden K. Sampson, Sepp Hochreiter, and Grey S. Nearing. Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resources Research*, 55(12): 11344–11354, 2019. doi: 10.1029/2019WR026065.
- [35] Grey Nearing et al. Global prediction of extreme floods in ungauged watersheds. *Nature*, 627:559–563, 2024. doi: 10.1038/s41586-024-07145-1.
- [36] Samuel Daramola, David F. Muñoz, and Chaopeng Shen. Toward transferable models for efficient spatiotemporal flood prediction across coastal-estuarine systems. *Cambridge Prisms: Coastal Futures*, 4:e13, 2026. doi: 10.1017/cft.2026.10037.
- [37] William H. Klein, Billy M. Lewis, and Isadore Enger. Objective prediction of five-day mean temperatures during winter. *Journal of Meteorology*, 16(6):672–682, 1959. doi: 10.1175/1520-0469(1959)016<0672:OPOFDM>2.0.CO;2.
- [38] Madeleine E. Harvey, Sarah N. Giddings, Geno Pawlak, and Jeffrey A. Crooks. Hydrodynamic variability of an intermittently closed estuary over interannual, seasonal, fortnightly, and tidal timescales. *Estuaries and Coasts*, 46(1):84–108, 2023. doi: 10.1007/s12237-021-01014-0.
- [39] Michael G. Hart-Davis, Gaia Piccioni, Denise Dettmering, Christian Schwatke, Marcello Passaro, and Florian Seitz. EOT20: a global ocean tide model from multi-mission satellite altimetry. *Earth System Science Data*, 13:3869–3884, 2021. doi: 10.5194/essd-13-3869-2021.
- [40] Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. The limits of

distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021. doi: 10.1093/imaiai/iaaa017.

- [41] Ivan D. Haigh, Marta Marcos, Stefan A. Talke, Philip L. Woodworth, John R. Hunter, Ben S. Hague, Arne Arns, Elizabeth Bradshaw, and Philip Thompson. GESLA version 3: A major update to the global higher-frequency sea-level dataset. *Geoscience Data Journal*, 10(3):293–314, 2023. doi: 10.1002/gdj3.174.
- [42] Ming Ji, Frank Aikman, and Carlos Lozano. Toward improved operational surge and inundation forecasts and coastal warnings. *Natural Hazards*, 53(1):195–203, 2010. doi: 10.1007/s11069-009-9414-z.
- [43] Muthukumaran Ramasubramanian and Brian J. Zikmund-Fisher. “When should I leave?”: The effect of impact-based hurricane wind and storm surge emergency alerts on evacuation intentions. *Weather, Climate, and Society*, 18(3):901–920, 2026. doi: 10.1175/WCAS-D-25-0215.1.
- [44] Christina Nauman, Eric Anderson, Erin Coughlan de Perez, Andrew Kruczkiewicz, Shanna McClain, Amanda Markert, Robert Griffin, and Pablo Suarez. Perspectives on flood forecast-based early action and opportunities for Earth observations. *Journal of Applied Remote Sensing*, 15(3):032002, 2021. doi: 10.1117/1.JRS.15.032002.
- [45] Andrew Kruczkiewicz, Zinta Zommers, Joyce Kimutai, Matthias Garschagen, Joshua Fisher, and Edson Severnini. The human and social impacts of climate overshoot. *PNAS Nexus*, 4(11):pgaf332, 2025. doi: 10.1093/pnasnexus/pgaf332.
- [46] Sanne Muis, Jeroen C. J. H. Aerts, José A. Á. Antolínez, Job C. Dullaart, Trang Minh Duong, Li Erikson, Rein J. Haarsma, Maialen Irazoqui Apecechea, Matthias Mengel, Dewi Le Bars, Andrea O’Neill, Roshanka Ranasinghe, Malcolm J. Roberts, Martin Verlaan, Philip J. Ward, and Kun Yan. Global projections of storm surges using high-resolution CMIP6 climate models. *Earth’s Future*, 11(9):e2023EF003479, 2023. doi: 10.1029/2023EF003479.
- [47] Andrew Kruczkiewicz, Richard Bernknopf, Yusuke Kuwayama, Humberto Vergara, Carolynne Hultquist, Fernanda Ayala, Juan Bazo, Bastian van den Bout, and Maarten van Aalst. Assessing the value of earth observations in the humanitarian sector: a framework to analyze the influence of EO within decision making. *Climate Risk Management*, 53:100857, 2026. doi: 10.1016/j.crm.2026.100857.
- [48] Susmita Dasgupta, Benoit Laplante, Siobhan Murray, and David Wheeler. Sea-level rise and storm surges: A comparative analysis of impacts in developing countries. Technical Report Policy Research Working Paper 4901, World Bank, Washington, DC, 2009.
- [49] Sven Kunze and Eric A. Strobl. The global long-term effects of storm surge flooding on human settlements in coastal areas. *Environmental Research Letters*, 19(2):024016, 2024. doi: 10.1088/1748-9326/ad18df.
- [50] Jason P. Dunion, Chris Davis, Helen Titley, Helen Greatrex, Munehiko Yamaguchi, John Methven, Raghavendra Ashrit, Zhuo Wang, Hui Yu, Anne-Claire Fontan, Alan Brammer, Matthew Kucas, Matthew Ford, Philippe Papin, Fernando Prates, Carla Mooney, Andrew Kruczkiewicz, Paromita Chakraborty, Andrew Burton, Mark DeMaria, Ryan Torn, and Jonathan L. Vigh. Recommendations for improved tropical cyclone formation and position probabilistic forecast products. *Tropical Cyclone Research and Review*, 12(4):241–258, 2023. doi: 10.1016/j.tcr.2023.11.003.
- [51] Emiliano Longo, Andrea Ficchi, Martin Verlaan, Sanne Muis, and Andrea Castelletti. A deep learning framework for extreme storm surge modeling under future climate scenarios. *Earth’s Future*, 14(3), 2026. doi: 10.1029/2025EF007072.
- [52] Alexander Y. Sun and Albert A. Sun. Zero-shot forecasting of streamflow using time series foundation models: are we there yet? *Machine Learning: Earth*, 2026. doi: 10.1088/3049-4753/ae4982.
- [53] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *AAAI*, 2023.
- [54] Daniel L. Codiga. Unified tidal analysis and prediction using the UTide Matlab functions. Technical Report 2011-01, Graduate School of Oceanography, University of Rhode Island, 2011.
- [55] Hans Hersbach, Bill Bell, Paul Berrisford, et al. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: 10.1002/qj.3803.
- [56] Frederik Kratzert, Daniel Klotz, Guy Shalev, Günter Klambauer, Sepp Hochreiter, and Grey Nearing. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, 23:5089–5110, 2019. doi: 10.5194/hess-23-5089-2019.

- [57] Olivier C. Pasche, Jonathan Wider, Zhongwei Zhang, Jakob Zscheischler, and Sebastian Engelke. Validating deep learning weather forecast models on recent high-impact extreme events. *Artificial Intelligence for the Earth Systems*, 4(1), 2025. doi: 10.1175/AIES-D-24-0033.1.
- [58] Zhongwei Zhang, Erich Fischer, Jakob Zscheischler, and Sebastian Engelke. Numerical models outperform ai weather forecasts of record-breaking extremes, 2025. arXiv:2508.15724.
- [59] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, et al. Chronos: Learning the language of time series, 2024. arXiv:2403.07815.

Extended Data

Extended Data Table 1: Season-scale real-forecast forcing check: the delivered model driven by GEFSv12 reforecast control forcing at every 00 UTC cycle of 2012 and 2017, against the same model with reanalysis (perfect-prognosis) forcing and against persistence, on identical continuous windows at the test gauges. Skills against persistence in percent from gauge-mean RMSE; storm-window RMSE over all leads inside p99.9 windows; capture = point forecast at the true peak hour as a fraction of peak height; q99 above peak = share of storm windows whose true peak lies under the q99 envelope. Where two or three values are given they refer to reanalysis forcing / reforecast forcing (/ persistence).

	2012	2017	both years
Gauges with continuous records	78	70	79
Forecast windows	18,432	16,796	35,228
Storm windows (p99.9)	141	142	283
Pooled skill, reanalysis / reforecast	+37.5 / +30.5	+37.8 / +30.2	+37.3 / +30.1
Skill at 8 h	+29.4 / +24.5	+27.9 / +24.9	+28.2 / +24.2
Skill at 48 h	+41.6 / +32.7	+42.2 / +33.6	+41.4 / +32.9
Gauges beating persistence, reforecast (reanalysis)	71/78 (73/78)	64/70 (66/70)	72/79 (74/79)
Storm-window RMSE, reanalysis / reforecast / persistence (cm)	12.1 / 14.3 / 24.8	14.5 / 16.4 / 24.7	13.4 / 15.4 / 24.7
Peak capture, reanalysis / reforecast (persistence)	0.66 / 0.69 (0.27)	0.65 / 0.69 (0.27)	0.65 / 0.69 (0.27)
q99 above the true peak, reanalysis / reforecast (%)	46 / 49	49 / 44	47 / 46

Extended Data Table 2: Rotating leave-region-out zero-shot transfer under the final protocol (validation-fold checkpoint selection, seismic-masked records, continuous evaluation windows; each region evaluated by a model trained without any data from it). Storm per-lead = skill range over leads 4 to 12 h inside p99.9 windows. The Oceania row is the median of three same-pipeline runs (pooled skill +5 to +10% across runs).

Withheld region	<i>n</i>	pooled	@8h	storm per-lead	peak capt. (pt / q99)	NNSE@8h
Japan + Pacific	77	+30%	+23%	+27 to +47%	0.59 / 0.93	0.803
Europe	231	+11%	+11%	+8 to +13%	0.38 / 0.69	0.719
North America	214	+10%	+9%	+17 to +20%	0.41 / 0.76	0.680
Oceania	111	+8%	+9%	−1 to +12%	0.43 / 0.75	0.727

Extended Data Table 3: Data scaling on the fixed 84-gauge marine test set (identical recipe at every size; the largest size is the delivered model; continuous evaluation windows). q99 env. = mean q99 height at the true peak hour as a fraction of mean peak height; tail cov. = q99 coverage of high-water hours.

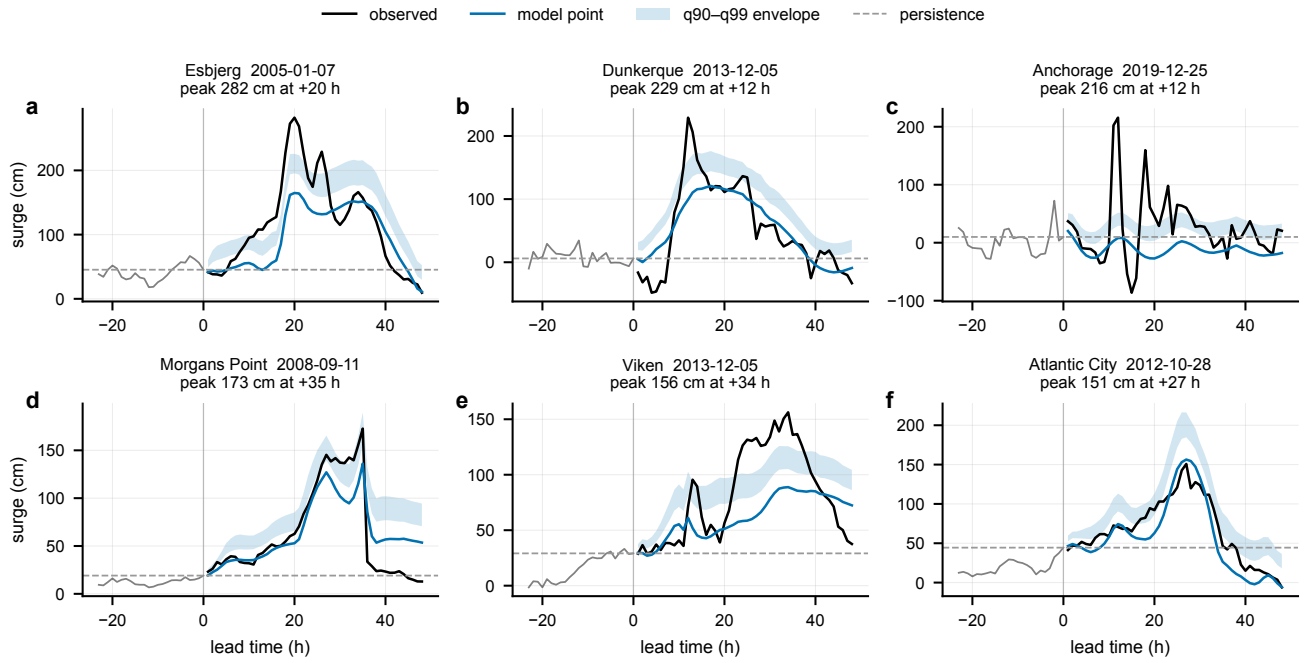
Training gauges	pooled	@8h	NNSE@8h	high-water p99.9	peak capt.	q99 env.	tail cov.
64	+19%	+17%	0.728	+26%	0.48	0.82	0.69
128	+24%	+20%	0.740	+32%	0.51	0.89	0.72
256	+28%	+23%	0.753	+34%	0.53	0.89	0.73
384	+31%	+27%	0.766	+39%	0.54	0.89	0.73
512	+35%	+28%	0.774	+43%	0.56	0.92	0.74
640	+36%	+30%	0.779	+45%	0.58	0.96	0.78
760	+37%	+30%	0.782	+45%	0.59	0.93	0.75

Extended Data Table 4: Controlled architecture factor study: 298-gauge training set, Japan withheld (77 test gauges), clean protocol (validation-fold checkpoint selection, seismic-masked records, 12 epochs for both trained rows), continuous evaluation windows. Skills against persistence; peak capture at the true peak hour of p99.9 windows (persistence 0.30). The recurrent model has the lower 8-h RMSE at 71 of 77 gauges against the hour- and channel-embedded attention model.

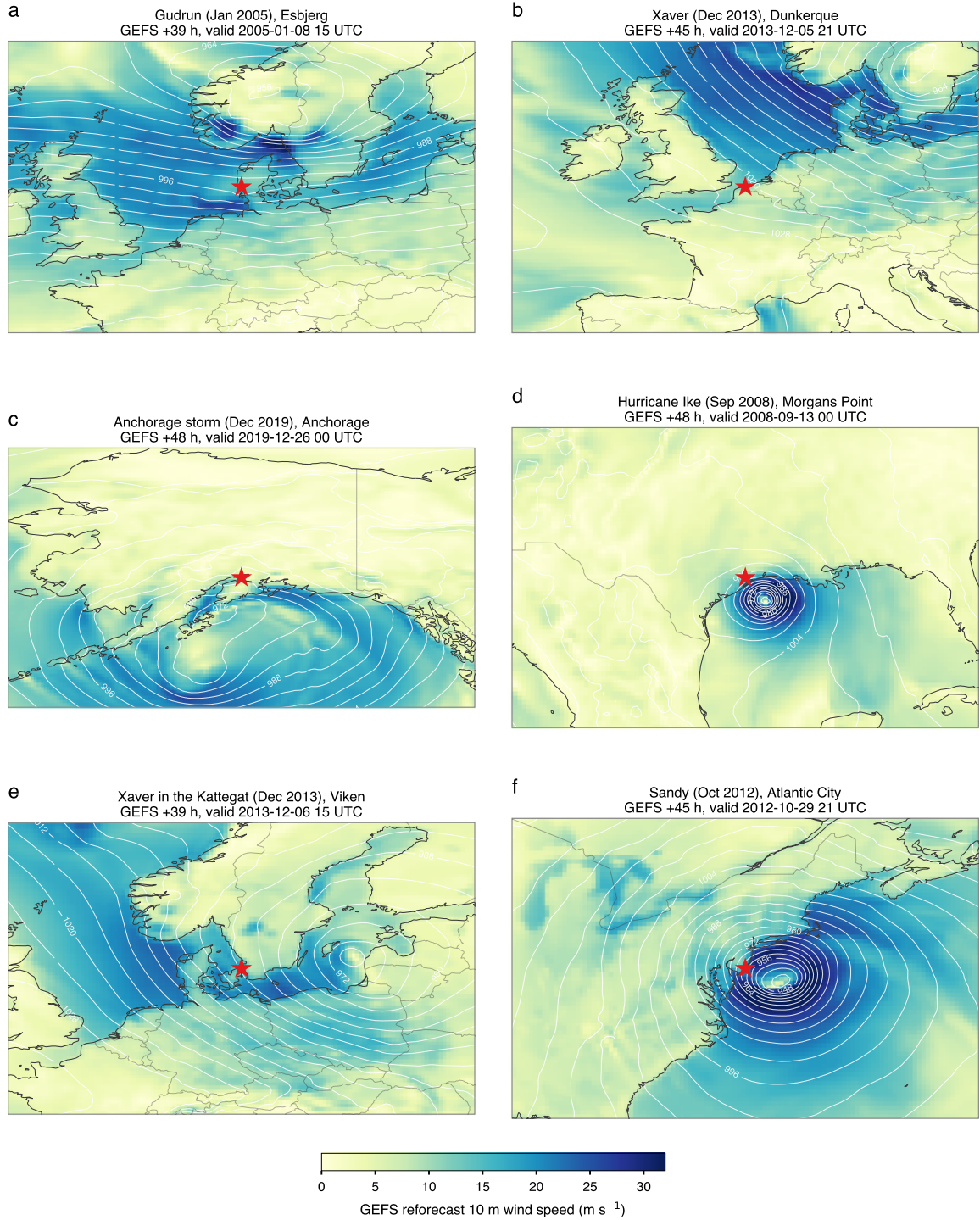
Model	backbone	forcing	pooled	@8h	high-water p99.9	peak capt.
Recurrent (delivered design)	recurrence	hourly, decoder	+15%	+14%	+31%	0.58
Attention, hourly forcing tokens with hour and channel embeddings	attention	hourly, cross-attention	+7%	0%	+11%	0.38
Chronos-bolt (generic pretrained, zero-shot)	attention	none	−1%	+5%	+3%	0.32
persistence			0	0	0	0.30

Extended Data Table 5: Every headline result on fully continuous evaluation windows (the primary set used throughout the paper) and on all evaluation windows. Windows are cut on the joined record after dropped hours are removed, so a fraction of them span a gap (the share that are 256 consecutive hours is given per row). Skills against persistence in percent; peak covered = share of p99.9 storm windows whose true peak lies under the q99 envelope. No comparison changes sign; one deployment gauge (Veracruz) is positive on all windows and negative on continuous windows (82 against 81 of 84 gauges beating persistence).

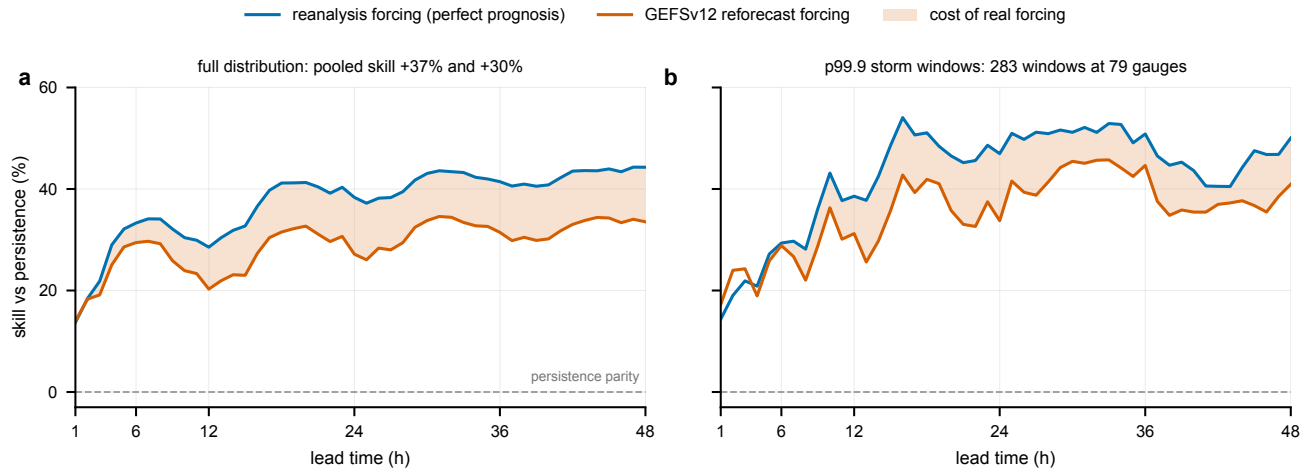
	pooled		@8h		NNSE@8h		high-water p99.9		peak capt.		peak covered %	
	cont.	all	cont.	all	cont.	all	cont.	all	cont.	all	cont.	all
Deployment, 84 gauges (71% continuous)	+37	+39	+30	+31	0.782	0.786	+45	+43	0.59	0.60	41	46
Rotation Japan (55%)	+30	+35	+23	+26	0.803	0.808	+34	+41	0.59	0.65	46	54
Rotation Europe (67%; 231 and 233 gauges)	+11	+13	+11	+12	0.719	0.719	+16	+16	0.38	0.39	23	26
Rotation North America (70%)	+10	+12	+9	+10	0.680	0.687	+20	+20	0.41	0.45	26	30
Rotation Oceania, median run (80%)	+8	+11	+9	+10	0.727	0.731	+10	+11	0.43	0.45	26	26
Europe intervention, 299 arm (67%)	0	+3	+7	+7	0.698	0.698	+4	+5	0.29	0.31	17	19
Ladder 64	+19	+21	+17	+17	0.728	0.731	+26	+28	0.48	0.50	34	38
Ladder 128	+24	+26	+20	+21	0.740	0.745	+32	+32	0.51	0.52	39	42
Ladder 256	+28	+30	+23	+23	0.753	0.756	+34	+34	0.53	0.54	40	44
Ladder 384	+31	+34	+27	+27	0.766	0.769	+39	+38	0.54	0.56	38	43
Ladder 512	+35	+37	+28	+29	0.774	0.778	+43	+40	0.56	0.57	41	46
Ladder 640	+36	+38	+30	+31	0.779	0.784	+45	+42	0.58	0.59	46	51
Composition 298, 49 countries	+29	+31	+24	+25	0.755	0.759	+39	+38	0.53	0.55	38	43
Composition 298, USA and Japan	+7	+12	−2	0	0.655	0.664	+6	+3	0.38	0.39	34	37



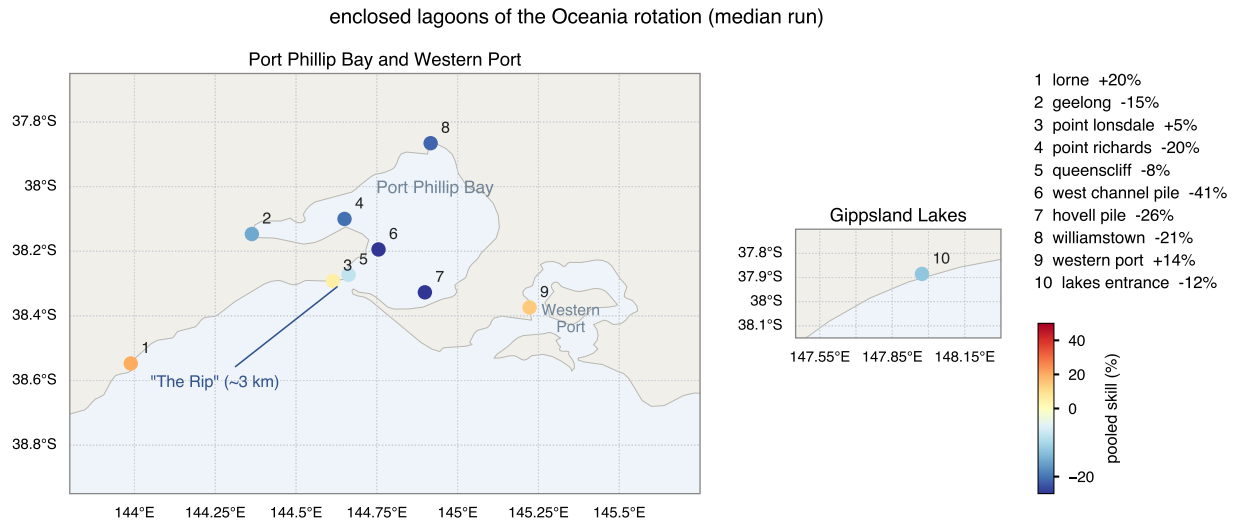
Extended Data Fig. 1: Largest extreme events across the 84 marine test gauges, forecast zero-shot by the delivered model: observed surge (black), point forecast (blue), q90 to q99 envelope (shaded), persistence (dashed). Panels (a) to (f): Gudrun (Esbjerg 2005, 2.8 m, peak at +20 h), Xaver (Dunkerque 2013), an Anchorage winter storm (2019; the diagnostic miss: Cook Inlet is a funnel-shaped macrotidal estuary where single-point forcing under-resolves both the wind fetch and the residual tidal dynamics, the same failure family as the bomb-cyclone and small-island cases), Hurricane Ike (Morgans Point 2008, 1.7 m at +35 h, rising limb and peak inside the envelope a day and a half ahead), Xaver in the Kattegat (Viken 2013), and Sandy (Atlantic City 2012, peak at +27 h). The selection lands, with no curation, on named historical storms.



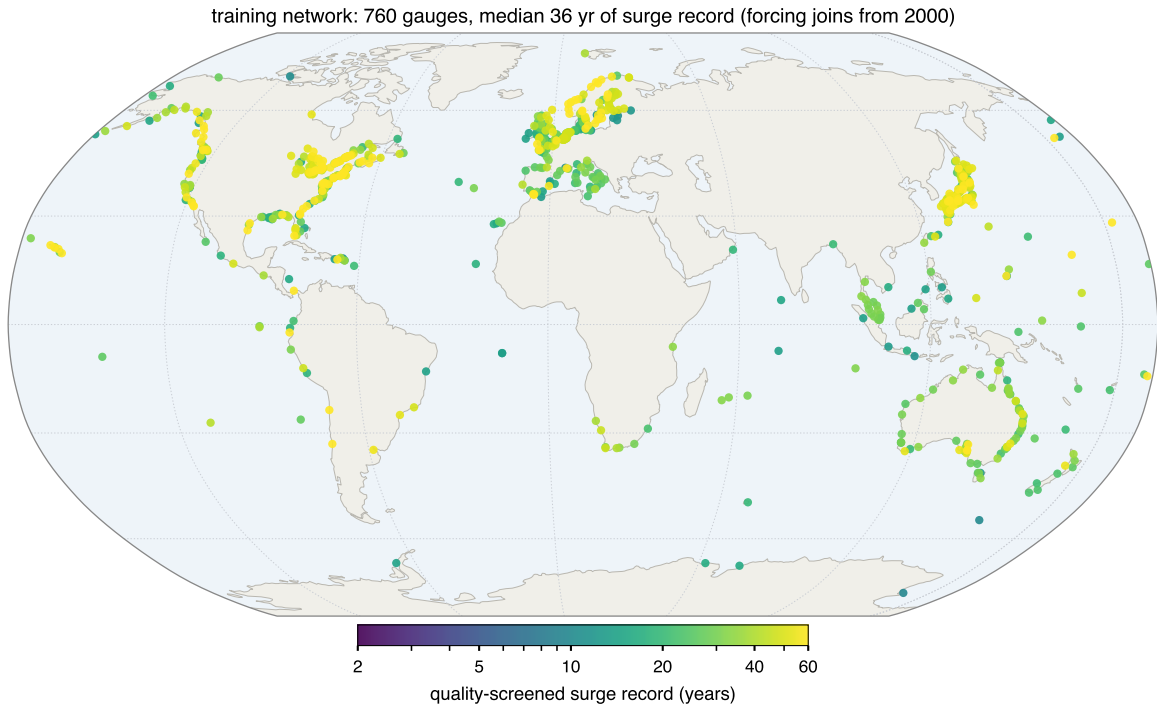
Extended Data Fig. 2: Storm context for Extended Data Fig. 1: GEFSv12 reforecast 10 m wind speed (shading) and mean sea-level pressure (contours, 4 hPa) at the forecast hour nearest each observed surge peak; the star marks the test gauge.



Extended Data Fig. 3: Skill against persistence by lead under reanalysis (perfect-prognosis) forcing and under GEFSv12 reforecast forcing, every 00 UTC cycle of 2012 and 2017 at 79 test gauges on identical continuous windows (Extended Data Table 1). (a) Full distribution; (b) p99.9 storm windows. Skill is computed from RMSE pooled over hours at each lead.

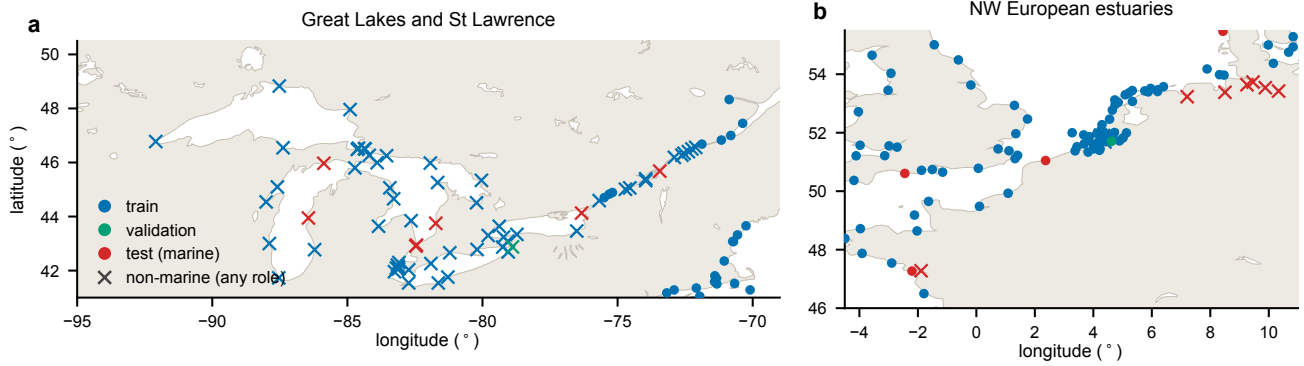


Extended Data Fig. 4: The regime limit up close. Gauges of the Oceania rotation (median run) around Port Phillip Bay, Western Port, and the Gippsland Lakes, numbered west to east with per-gauge pooled skill listed at right. Open-coast gauges (Lorne, Point Lonsdale, Western Port) are positive; every gauge inside the narrow-mouthed enclosures is negative in all three clean runs.



Extended Data Fig. 5: The training network: 760 gauges coloured by quality-screened surge record length (median 36 yr; the forcing join begins in 2000). Europe and Japan are dense with half-century records, while Africa and South America contribute few and short records, which is also where zero-shot forecasts have the fewest regime analogues to draw on.

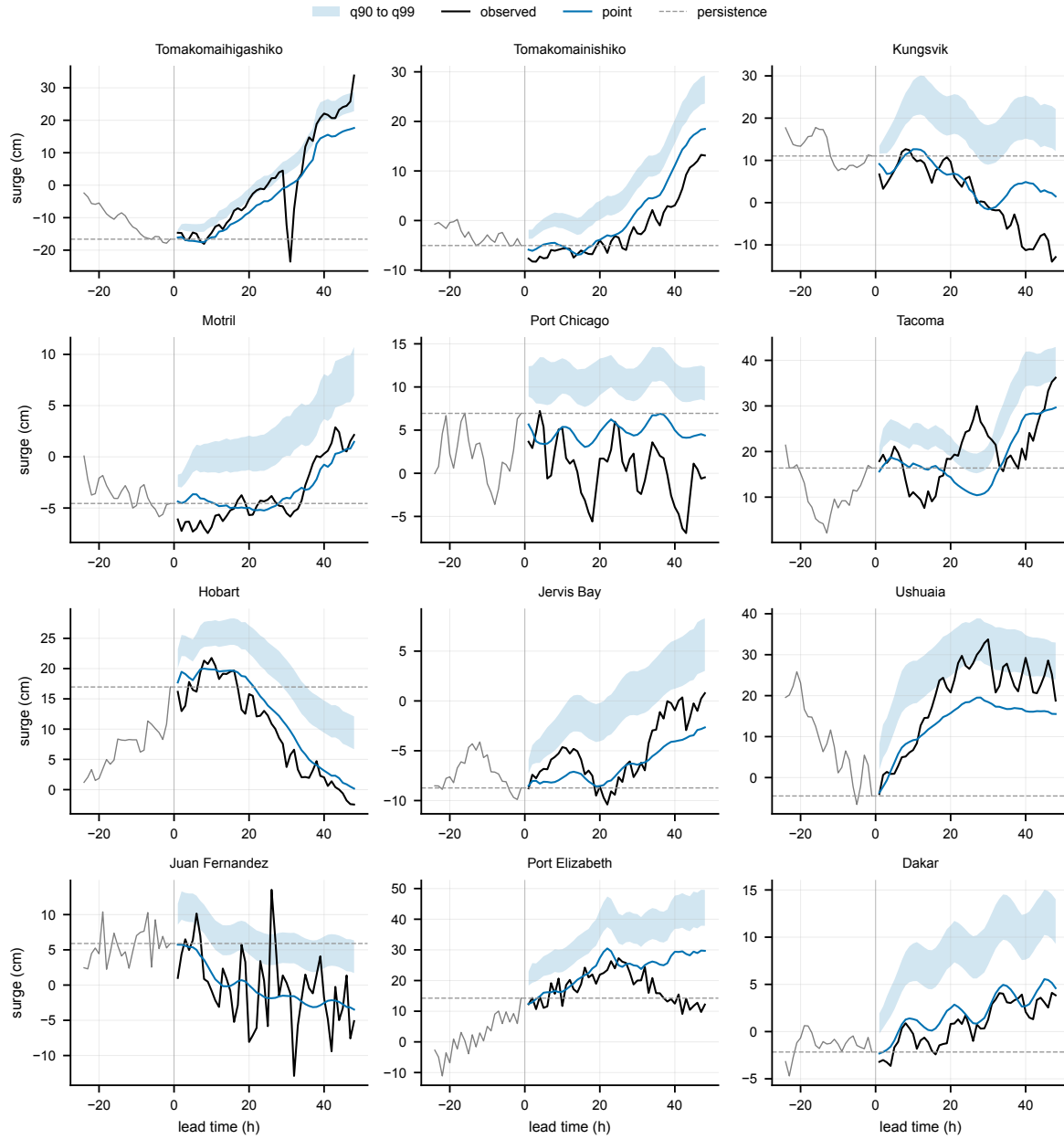
Supplementary Information



Supplementary Fig. 1: The two regions where non-marine gauges cluster, with the same encoding as Fig. 1a plus domain markers (crosses = lake, river, or estuary gauges; colour gives the role). (a) Great Lakes and St Lawrence: crosses on the river are the fluvial reach above the head of tide at Portneuf (tidal range 0.01 to 1.9 m), including the primary test gauge Varennes, which stays in the primary set as registered; the uncrossed gauges at and seaward of Portneuf are the macrotidal lower estuary (ranges 4.0 to 6.0 m) and are retained as coastal. (b) Northwest European estuaries: the Elbe, Ems, and Loire chains, the Rhine–Meuse tidal rivers, and one Dutch inland waterway. Of the 62 crossed training gauges, the 54 in the Great Lakes, on the St Lawrence from Lac Saint-Pierre to Deschaillons, and on the Dutch inland waterway were removed in the marine-only ablation (Methods); the four between Lake Ontario and Montreal and the four Rhine–Meuse gauges were not. Three isolated non-marine sites are not shown: the Sacramento delta and Columbia River gauges (out-of-domain test) and the Antarctic ice-shelf station.

Out-of-domain gauges. The sixteen pre-registered gauges relabelled out of domain (five Great-Lakes, one Antarctic, ten estuary, river, weir, or lake sites) were evaluated separately with the delivered model (15 of 16 pass the joined-hours minimum; all windows). The model beats persistence on 14 of 15 (pooled +37%, @8h +34%, storm per-lead +32 to +42%, q99 envelope 0.88), confirming that their exclusion from the primary set reflects domain definition, not performance.

1435



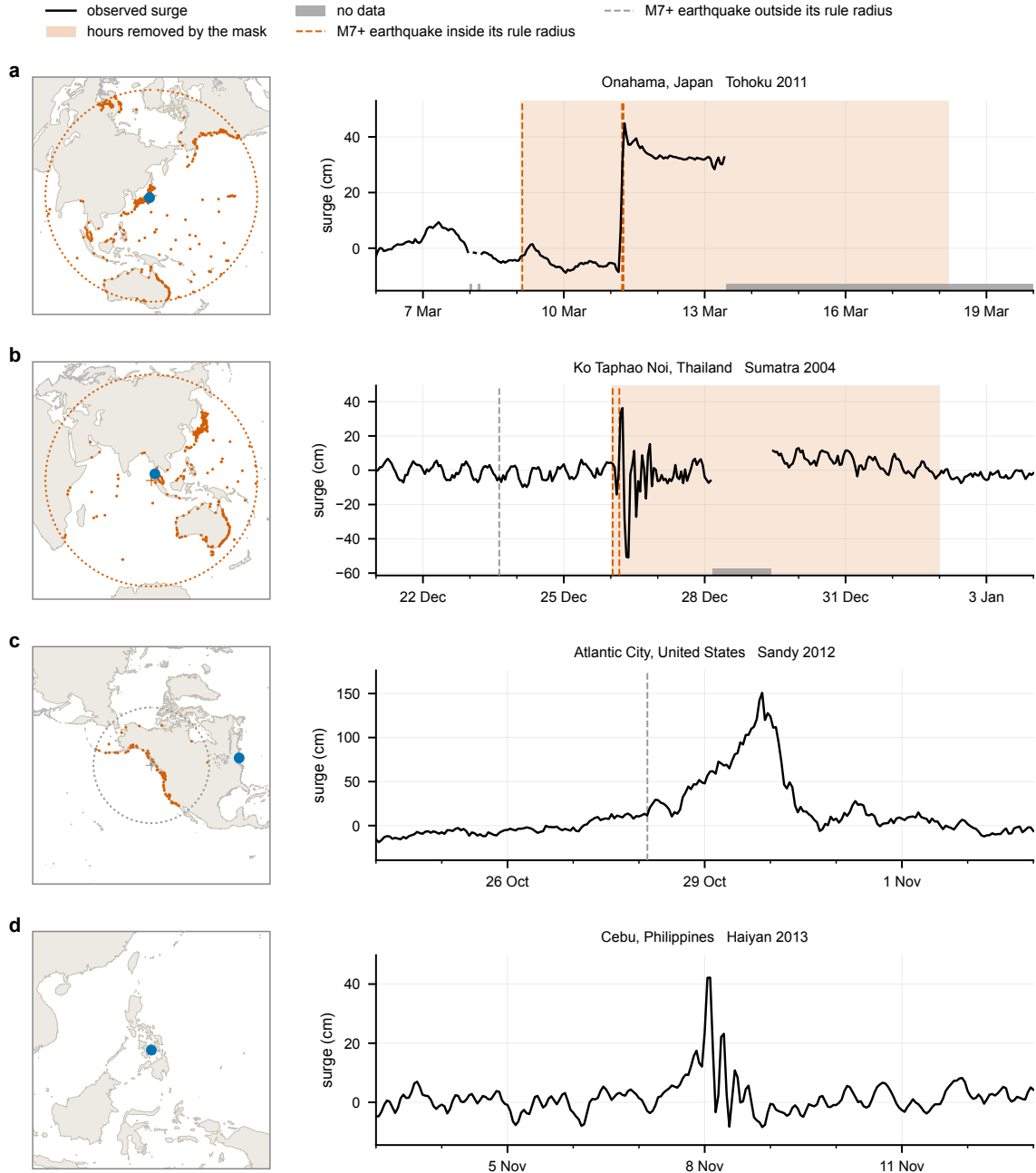
Supplementary Fig. 2: Twelve unselected forecast windows: random marine test gauges (continent-stratified, seed 0), random issue times, no selection criterion. Observed surge (black), point forecast (blue), q90 to q99 envelope (shaded), persistence (dashed), 24 h of context left of the vertical line. Envelopes stay narrow in calm conditions, moderate events are tracked, and the failure modes (noisy semi-enclosed harbours, remote islands) are visible.

Supplementary Table 1: Reported metrics, definitions, and implementation sections of `scripts/eval_full.py`. All statistics are computed per gauge and averaged across gauges unless noted (high-water-hour skill pools hours across gauges and is therefore window-count-weighted); all thresholds (p99, p99.5, p99.9) are per-gauge percentiles of the hourly surge distribution over that gauge’s full test record.

Metric	Definition	Script section
Per-lead RMSE / MAE / skill	Eq. (5); skill from gauge-averaged model against persistence RMSE	[A]
Pooled RMSE / skill	Eq. (6) over $\ell \leq 48$, then as in Eq. (5)	[A]
NNSE	$1/(2 - \text{NSE})$, per gauge at fixed lead, averaged	[A]
Event-window RMSE	RMSE restricted to 48-h windows containing surge $\geq$ threshold	[B]
High-water-hour RMSE	RMSE restricted to hours with $y \geq$ threshold	[B]
Peak bias / MAE / capture	At true peak hour of each p99.9 window; capture per Eq. (7)	[B]
Per-lead skill in events	Eq. (5) on squared errors pooled over all p99.9 windows across gauges	[B]
Peak capture by lead position	Eq. (7) stratified by peak lead bucket (1–6, 7–12, 13–24, 25–48 h)	[C]
Coverage (overall / tail)	Eq. (8)	[D]
Sharpness ratio	$\langle \hat{y}^{(q99)} - \hat{y}^{(\text{pt})} \rangle$ storm hours / calm hours	[D]
Crossing rate	$\Pr(\hat{y}^{(q90)} > \hat{y}^{(q99)})$	[D]
Uncertainty	Spatial-block bootstrap over eleven ocean regions (4,000 resamples; Methods)	post-hoc
Window continuity	Windows whose 256 rows are consecutive hours (primary set); all windows as sensitivity	[W]

Supplementary Table 2: Storm-case real-forecast check: the same model driven by GEFsV12 reforecast control forcing, against reanalysis (perfect-prognosis) forcing and persistence, on identical 48-h windows for the nine largest test-gauge storm cases with complete coverage (eight storms; Xavier at two gauges). Issue = the 00 UTC reforecast cycle 24 to 48 h before the catalogued storm peak; lead = hours from issue to the catalogued peak; peak = catalogued peak height (inside the window for every storm). RMSE over the 48-h window in cm; capture = point forecast at the true peak hour as a fraction of peak height; q99 = envelope height at the true peak hour as a fraction of peak height (a value below 1 means the envelope lies under the peak).

Storm (gauge)	issue	lead h	peak cm	window RMSE			peak capture		q99 / peak		
				rean.	GEFS	pers.	rean.	GEFS	rean.	GEFS	
Gudrun 2005 (Esbjerg)	2005-01-07	40	282	52	57	102	0.50	0.37	0.71	0.55	
Xaver 2013 (Dunkerque)	2013-12-04	45	229	28	34	68	0.49	0.50	0.74	0.77	
Ike 2008 (Morgans Point)	2008-09-12	29	173	18	63	104	0.79	0.80	1.08	1.12	
Xaver 2013 (Viken)	2013-12-05	40	156	36	64	69	0.59	0.19	0.82	0.59	
Sandy 2012 (Atlantic City)	2012-10-28	45	151	11	21	63	0.89	0.65	1.26	1.10	
Anchorage 2010/11	2010-12-31	31	142	47	49	50	0.05	−0.01	0.27	0.26	
Sandakan 2009	2009-12-14	43	141	56	57	105	0.18	0.17	0.42	0.40	
Ballen 2012	2012-02-08	40	131	75	77	76	−0.13	−0.16	0.09	0.05	
Saint-Nazaire 2017	2017-12-10	27	127	10	22	26	0.90	0.71	1.15	0.94	
mean				37.1	49.3	73.8	0.47	0.36	0.73	0.64	



Supplementary Fig. 3: Seismic mask spot checks. Each row is one event. Left: azimuthal equidistant map centred on the largest M7+ earthquake in the window (cross; orange where the event lies within the rule radius of its magnitude class from the example gauge, grey otherwise), the rule radius of that class (dotted circle), the split gauges inside it (orange dots), the other split gauges (grey dots) and the example gauge (blue dot). Right: hourly surge at the example gauge; shading marks the hours the mask removes, dashed lines mark every M7+ earthquake in the window with the colours of the map, and the grey bar along the axis marks hours with no data. (a) Tohoku 2011 at Onahama, Japan: the M9.1 mainshock 200 km away triggers the 8,000 km, 168-h class; the tsunami signal (45 cm) is removed, 105 of the 175 recorded hours are removed (the record has no rows after 13 March 10:00 UTC) and 354 gauges lie inside the radius. (b) Sumatra 2004 at Ko Taphao Noi, Thailand: the M9.1 mainshock 572 km away removes the tsunami oscillation (−51 cm; 138 h removed; 276 gauges inside the radius). (c) Sandy 2012 at Atlantic City, United States: the M7.8 Haida Gwaii earthquake lies 4,556 km away, outside the 3,000 km radius of its class, so the 151 cm storm peak is retained while 76 Pacific gauges lose hours. (d) Haiyan 2013 at Cebu, Philippines: no M7+ earthquake in the window; the 42 cm peak is retained. The rule is stated in Methods; Supplementary Table 3 tabulates the same events and the gauges the mask affects most.

Supplementary Table 3: Seismic mask: behaviour on four events and the gauges it affects most. The mask removes hours within graded windows of USGS M7+ earthquakes ($M \geq 7.0$ within 1,000 km: 72 h; $M \geq 7.5$ within 3,000 km: 96 h; $M \geq 8.0$ within 8,000 km: 168 h; Methods). (a) Two tsunamis the mask removes and two storms it leaves intact. Earthquakes: every M7+ event in the window with its distance to the example gauge; events inside the rule radius of their magnitude class are in bold. Strongest tier: the magnitude class, radius and duration of the longest window triggered at the gauge. Hours removed: rows of the gauge’s record inside the mask windows; the count falls below the nominal span where the record has gaps (Ko Taphao Noi lacks 31 of the 337 hours in its window; the Onahama record has no rows after 13 March 10:00 UTC in this window, so the rule spans 220 hours but only 175 rows exist, of which 105 are removed). Gauges affected: the number of the 900 split gauges inside the rule radius of the largest earthquake in the window. Largest surge: the largest absolute surge in the window at the example gauge and whether the mask removes it. During Sandy an M7.8 earthquake off Haida Gwaii lay 4,556 km from Atlantic City, outside the 3,000 km radius of its class, so no hour was removed there while 76 Pacific gauges lost hours; no M7+ earthquake occurred during Haiyan. (b) The sixteen gauges that lose the largest share of their p99.9 exceedance hours to the mask, with the share and the number of exceedance hours.

a

Event	Example gauge (fold)	M7+ earthquakes in the window (distance to the gauge)	Strongest tier triggered at the gauge	Hours removed	Gauges affected	Largest surge in the window
Tohoku 2011	Onahama, Japan (test)	9 Mar 02:45 M7.3, 239 km; 11 Mar 05:46 M9.1, 200 km; 11 Mar 06:15 M7.9, 76 km; 11 Mar 06:25 M7.7, 349 km	M8.0+: 8,000 km, 168 h	105	354	45 cm, removed
Sumatra 2004	Ko Taphao Noi, Thailand (train)	23 Dec 14:59 M8.1, 8,785 km; 26 Dec 00:58 M9.1, 572 km; 26 Dec 04:21 M7.2, 612 km	M8.0+: 8,000 km, 168 h	138	276	−51 cm, removed
Sandy 2012	Atlantic City, United States (test)	28 Oct 03:04 M7.8, 4,556 km	none	0	76	151 cm, kept
Haiyan 2013	Cebu, Philippines (train)	none	none	0	0	42 cm, kept

b

Gauge	Fold	p99.9 exceedance hours removed	Exceedance hours
Vung Tau, Vietnam	train	45%	145
Veracruz, Mexico	test	25%	142
Rough and Ready Island, United States	train	20%	177
Port Vila, Vanuatu	train	20%	219
Nouméa, New Caledonia (France)	test	15%	129
Pago Bay, Guam (United States)	train	13%	129
Nagasaki, Japan	test	13%	298
Kahului, United States	train	13%	552
Pointe-à-Pitre, Guadeloupe (France)	train	12%	180
Trident Pier, United States	train	12%	232
Port Chalmers, New Zealand	train	11%	203
Onahama, Japan	test	11%	459
Cabo San Lucas, Mexico	train	11%	216
Esashiko, Japan	train	11%	304
Virginia Key, United States	train	10%	239
Daugavgriva, Latvia	train	10%	107