https://doi.org/10.5194/egusphere-2026-5073
Preprint. Discussion started: 31 August 2026

MAROR: a multipoint autoregressive network for forcing-driven, uncertainty-calibrated reconstruction of hourly coastal sea level from 1940 to the present

Angel Amores[1,2], Roberto Alcaraz[1], Marta Marcos[1,2], Philip Thompson[3], Ariadna Martín[4,5], and Thomas Wahl[4,5]

[1]Instituto Mediterráneo de Estudios Avanzados (UIB-CSIC), Esporles, Spain
[2]Departament de Física (UIB), Palma, Spain
[3]Department of Oceanography, University of Hawai'i at Manoa, Honolulu, HI 96822, USA
[4]Department of Civil, Environmental and Construction Engineering & National Central for Integrated Coastal Research, University of Central Florida, Orlando, USA
[5]National Center for Integrated Coastal Research, University of Central Florida, Orlando, FL, 32816, USA

Correspondence: Angel Amores (angel.amores@uib.es)

Abstract. Tide gauge records are the primary source for coastal sea level research, yet they are short, gappy, and affected by instrumental errors, which limits the study of extreme events and their long-term changes. We present MAROR (**M**ultipoint **A**utoregressive **R**econstruction **O**f sea-level **R**ecords), a neural network trained on tide gauge records that reconstructs gap-free, hourly coastal sea level from 1940 to the present, driven only by atmospheric forcing and, when available, satellite altimetry. Building on the HIDRA3, a convolutional neural network designed for short-term predictions, MAROR turns a short-range forecast model into a long-range autoregressive reconstruction engine, with an exposure-bias correction that removes drift and a generative ensemble that yields a calibrated, per-tide-gauge uncertainty band. We evaluate it on 373 tide gauges across eleven regions spanning a wide range of surge regimes, from extratropical shelves to tropical cyclones and western boundary currents. MAROR reconstructs the hourly storm surge with a median correlation of 0.88, reproduces the most extreme events, and outperforms both a hydrodynamic hindcast and a state-of-the-art deep-learning model.

1 Introduction

Tide gauges are the primary source of coastal sea level measurements and have been monitoring coastal sea level changes for decades and, at some locations, even for over a century. Nowadays, tide gauges have become a widespread technology, enabling the monitoring of large stretches of coastal regions worldwide (Haigh et al., 2023). These records, together with satellite altimetry observations, are essential to understand the mean and extreme sea level variability which affects coastal assets and the livelihood of the millions of people living in coastal regions. They are also relevant to oceanographic, hydrological and geodetic studies, among others, to the extent that most of the scientific research on sea level changes relies on tide gauge data, alone or in combination with satellite altimetry.

https://doi.org/10.5194/egusphere-2026-5073
Preprint. Discussion started: 31 August 2026

Sea level time series from tide gauges are generally provided at high sampling rates, typically hourly or finer. Tide gauge operators and data centres often apply different levels and types of quality control, inspecting the records for outliers and shifts and flagging suspicious values. However, tide gauge time series are rarely free of errors and gaps, particularly when decadal-long records are concerned. Datum shifts can occur due to unreported or uncontrolled changes in the instrumentation or the position of the instrument. Likewise, time shifts due to malfunctioning of the tide gauge clock can often be detected with a close inspection of the records. Outliers, in the form of isolated spikes, are also common. This implies that, for sea level records to be usable in research studies, they must be carefully inspected. For example, when the atmospherically induced contribution to sea level (that is, the storm surge) is computed by subtracting the astronomical tide from the observed sea level, time shifts become prominent signals that must be either corrected or discarded in the subsequent analyses. All in all, the manual quality control of tide gauge data is necessary but also time consuming, and it often involves subjective choices.

Despite the relatively good tide gauge coverage of many coastal regions, most of the sea level records are relatively short, spanning only a few decades or less (Haigh et al., 2023). This limits their use to investigate long-term changes in sea level, particularly those related to changes in extreme sea level events. Extremes are rare events, so that long, at least multi-decadal, records are required to characterise them and to quantify possible long-term changes and trends in response to climate drivers, including global warming. The usual approaches to extend the coastal sea level records in time are based on modelling. Hydrodynamic models solve the shallow water equations over a regional or global ocean domain forced by atmospheric pressure, surface winds and tidal oscillations, resulting in homogeneous, gap-free time series along the coasts (Agulles et al., 2024; Muis et al., 2020). Data-driven models, instead, learn a statistical relationship between forcing variables, such as atmospheric pressure, surface wind and sea surface temperature, and the sea level observed at a tide gauge, ranging from classic regression approaches (Cid et al., 2017; Tadesse et al., 2020) to modern machine learning (Tiggeloven et al., 2021; Nagaraj et al., 2025). Both families extend the record over the period spanned by the forcing, and their strengths are complementary: the hydrodynamic approach is physically consistent and spatially complete, whereas the data-driven one is based on the local observations and can capture site-specific behaviour that a model grid does not resolve.

Among the data-driven methods, those based on deep learning have proven especially powerful, owing to the ability of neural networks to capture the nonlinear relationships between observed phenomena and their physical drivers. They offer efficient and robust predictions of geophysical variables and have been successfully applied to weather forecasting (Bi et al., 2023; Lam et al., 2023), seasonal prediction (Ham et al., 2019) and extreme events (Camps-Valls et al., 2025; Materia et al., 2024). In the coastal sea level context, a deep-learning model (HIDRA) was developed to forecast high-frequency sea level oscillations in the Adriatic Sea from tidal and atmospheric forcing (Žust et al., 2021; Rus et al., 2023, 2025). Configured to run in operational mode, it captures sea level oscillations accurately, including the extremes, and largely outperforms the data-driven methods available so far. These models are, however, conceived as short-range forecasting tools, predicting a few days ahead rather than reconstructing the historical record.

In this work we build on the HIDRA3 architecture (Rus et al., 2025) and turn it from a short-range forecast model into a reconstruction engine, which we name MAROR (**M**ultipoint **A**utoregressive **R**econstruction **O**f sea-level **R**ecords), able to rebuild gap-free hourly coastal sea level from 1940 to the present. Three developments make this possible: a long-range

https://doi.org/10.5194/egusphere-2026-5073
Preprint. Discussion started: 31 August 2026

autoregressive rollout that reconstructs decades rather than a short forecast horizon, corrected for the drift that such rollouts would otherwise accumulate; the ingestion of satellite altimetry, which lets the network capture the slow ocean variability that the atmospheric forcing alone does not contain; and a generative ensemble that provides a calibrated, per-tide-gauge and per-lead-time uncertainty band, so that every reconstructed value carries its own confidence estimate. We train and evaluate the model on 373 high-frequency tide gauge records distributed worldwide, covering coastal regimes as different as extratropical storm surges on wide continental shelves, tropical cyclones, and settings dominated by western boundary currents, all of which have undergone an accurate and careful quality control so that the training relies on the most reliable data available. We show that MAROR reproduces the observed sea level variability down to the most extreme events, that it fills data gaps and extends sea level records backwards in time, and that, being physically plausible by construction, it can be used as an independent diagnostic to flag defects in the observations themselves.

2 The MAROR model

2.1 Overview

MAROR is a neural network that reconstructs gap-free hourly time series of total coastal sea level (astronomical tide plus the meteorologically and oceanographically driven components) at a set of tide gauges. It is based on the HIDRA3 architecture (Rus et al., 2025) further extended in three important aspects: (i) a long-range *autoregressive* rollout that reconstructs decades rather than a short forecast horizon, with an exposure-bias correction that removes the drift such rollouts otherwise accumulate; (ii) satellite altimetry (sea-level anomaly, SLA) ingested so that the network can capture the slow ocean variability that atmospheric reanalysis alone misses; and (iii) a generative ensemble that yields a per-gauge, per-lead-time uncertainty band, calibrated by a light post-hoc inflation.

The model is trained as a one-hour predictor: given the recent sea level at each tide gauge, the known astronomical tide, a window of atmospheric forcing over the region, and (optionally) the local SLA, it generates the sea level one hour ahead at every tide gauge within the region simultaneously. Applying this predictor recursively, i.e. feeding each prediction back in as the most recent input, turns it into a reconstruction engine that fills arbitrarily long gaps, including the historical storm surges, for hours at which no observation exists. The reconstruction is run *fully self-fed*: it is cold-started from the astronomical tide plus the inverse barometer and thereafter consumes only its own predictions, so no observed sea level is ever ingested; it is therefore driven by the atmospheric forcing and the tide alone, which is what lets it run continuously back to the beginning of the atmospheric forcing period, even at tide gauges and periods where observations are sparse or absent. Being a learned, forcing-driven emulator that delivers a calibrated per-gauge, per-lead-time uncertainty is what distinguishes MAROR from purely hydrodynamic hindcasts such as GTSM-ERA5 and the CODEC/GTSR reanalyses (Muis et al., 2016, 2020).

https://doi.org/10.5194/egusphere-2026-5073
Preprint. Discussion started: 31 August 2026

2.2 Inputs

Let N denote the number of tide gauges belonging to a region (the network is trained separately per region). At each prediction hour the model receives four inputs:

- **Sea-level history** $\mathbf{h} \in \mathbb{R}^{N\times 72}$: the last 72 h of total sea level at each tide gauge. In training this is the observed sea level (teacher forcing); in reconstruction it is filled entirely by the model's own predictions meaning that the record is reconstructed fully self-fed from a tide-plus-inverse-barometer cold start, so no observed sea level is ever used as input.
- **Astronomical tide** $\mathbf{t} \in \mathbb{R}^{N\times 144}$: the harmonic tide from 72 h in the past to 72 h in the future. The tide is deterministic and known ahead of time, so the future half of the window is a legitimate input and supplies the network with the tidal phase it is predicting into.
- **Atmospheric and oceanic forcing** $\mathbf{G} \in \mathbb{R}^{144\times 8\times H\times W}$: a 144 h window (again $\pm$72 h) of atmospheric and oceanic fields on the regional grid of size $H \times W$. The grid size $H \times W$ is region-specific (each region covers a domain of a different extent) and, because the geophysical encoder collapses any $H \times W$ grid to a single feature vector (Sect. 2.3.1), the same network applies unchanged across regions and the grid can be chosen freely for each one. The atmospheric and oceanic fields are, in our case, obtained from ERA5 reanalysis (Hersbach et al., 2020). They are downsampled from their native $0.25°$ resolution to $1°$, as in HIDRA3; tests at $0.5°$ brought no significant improvement, so $1°$ spatial resolution is retained. The eight channels are, following HIDRA3 architecture, mean sea-level pressure; the two 10 m wind components; sea-surface temperature; and four wave parameters (mean wave period, significant wave height, and the sine and cosine of the mean wave direction). Channels are standardised per field over ocean cells; land cells are set to zero (the standardised mean), a neutral fill.
- **Sea-level anomaly** $\boldsymbol{\ell} \in \mathbb{R}^{N\times 30}$ (optional): a 30-day window of daily satellite altimetry co-located with each tide gauge. SLA is only available from 1993 onward; where it is absent it is set to zero and, as described below, the network is initialised so that it starts identical to the no-SLA configuration.

2.3 Network architecture

MAROR proceeds through five modules (Fig. 1). Geophysical variables (wind, pressure, SST and waves) are encoded once for all tide gauges within the selected region in the *geophysical encoder* (Sect. 2.3.1), while the sea-level history, tide and SLA are processed separately for each gauge in the *feature extraction* module (Sect. 2.3.2). The per-gauge features are merged into a single joint state vector in the *feature fusion* module (Sect. 2.3.3), from which the *sea-level regression* module (Sect. 2.3.4) reads out the point estimate at each tide gauge. Finally, the *generative ensemble* (Sect. 2.3.5) turns latent noise into an ensemble of realisations whose spread is the predictive uncertainty. The first four modules descend from HIDRA3 (Rus et al., 2025); the methodological additions of MAROR are the one-hour autoregressive regression, the SLA backbone input, and the generative ensemble in place of the Gaussian variance head.

https://doi.org/10.5194/egusphere-2026-5073
Preprint. Discussion started: 31 August 2026

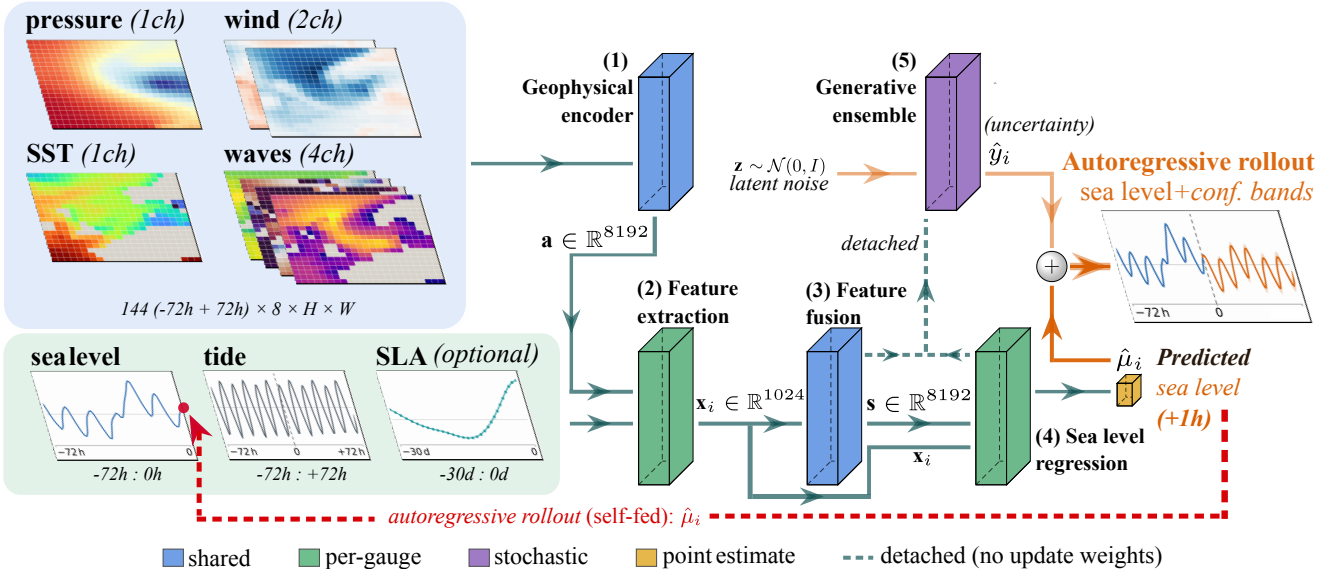

Figure 1. Overview of the MAROR architecture. The regional forcing fields (top left) and the per-gauge inputs (bottom left: sea level, astronomical tide and, optionally, satellite altimetry) feed the five modules of the network: the geophysical encoder, the feature extraction, the feature fusion, the sea-level regression and the generative ensemble. The regression module returns the sea level one hour ahead at every tide gauge; fed back as input (dashed red), it becomes the self-fed autoregressive rollout that reconstructs decades of hourly sea level together with its confidence bands (right). Colours identify the shared, per-gauge and stochastic parts of the network, as given in the legend.

2.3.1 Geophysical encoder

The geophysical encoder compresses the forcing window into a single feature vector shared by all tide gauges in the same region, and is inherited unchanged from HIDRA3 (Rus et al., 2025); we describe it here so that the model is self-contained. The eight forcing channels are grouped into four variables by physical role: wind $\mathbf{v} = (u_{10}, v_{10}) \in \mathbb{R}^{2\times144\times H\times W}$, pressure $\mathbf{p} \in \mathbb{R}^{1\times144\times H\times W}$, sea-surface temperature $\mathbf{T} \in \mathbb{R}^{1\times144\times H\times W}$, and waves $\mathbf{w} \in \mathbb{R}^{4\times144\times H\times W}$ (the sine and cosine of the mean wave direction, the mean wave period and the significant wave height), where $H \times W$ is the size of the region grid. Each variable is encoded independently, so that the encoder models each field on its own rather than their low-level cross-interactions, which keeps the parameter count down.

The encoding proceeds in two steps. In the first step, a variable of size $c \times 144 \times H \times W$ (c is that variable's number of channels: 1 for pressure and SST, 2 for wind, 4 for waves) is processed by a three-dimensional convolution with 64 kernels of size $2 \times 3 \times 3$ and stride $2 \times 2 \times 2$ (the first dimension is temporal, the other two spatial), a rectified linear unit (ReLU), three-dimensional dropout, and a second three-dimensional convolution with kernels of size $2 \times H_1 \times W_1$ and stride $2 \times 1 \times 1$, where $H_1 = \lfloor (H-3)/2 \rfloor + 1$ and $W_1 = \lfloor (W-3)/2 \rfloor + 1$ are the spatial dimensions after the first stride. The spatial kernel of the second convolution therefore matches the reduced grid exactly, collapsing it to 1×1 for *any* region grid (the property that lets the same encoder serve every region unchanged) while the two temporal strides reduce the 144 h window to 36 steps. The number of output kernels of the second convolution is 512 for wind and pressure but only 64 for SST and waves: the fields

https://doi.org/10.5194/egusphere-2026-5073
Preprint. Discussion started: 31 August 2026

that carry the most surge-relevant information are given the wider representation, following HIDRA3. Concatenating the four encodings along the channel axis gives $2 \cdot 512 + 2 \cdot 64 = 1152$ channels over 36 time steps.

In the second step, this 1152×36 tensor is mixed in time by a one-dimensional convolution with 256 kernels of temporal size 5 (chosen so its receptive field spans roughly 20 h), reducing the temporal dimension to 32, followed by two one-dimensional convolutions with 256 kernels of temporal size 1, each with a SELU activation (Klambauer et al., 2017), dropout and a residual connection. Flattening the result yields a single geophysical feature vector $\mathbf{a} \in \mathbb{R}^{8192}$ that summarises the regional atmospheric and wave state and is shared by all per-gauge encoders.

2.3.2 Feature extraction

At any prediction hour some tide gauges may lack a usable input window because of a data gap. The feature extraction module therefore processes each tide gauge independently, producing a location-specific feature vector $\mathbf{x}_i$ that the fusion module (Sect. 2.3.3) later merges across gauges. This module follows HIDRA3, with one addition: SLA enters here as a per-gauge backbone input.

For gauge i, the sea-level history $\mathbf{h}_i \in \mathbb{R}^{72}$, the astronomical tide $\mathbf{t}_i \in \mathbb{R}^{144}$ and, when SLA is enabled, the altimetry window $\boldsymbol{\ell}_i \in \mathbb{R}^{30}$ are concatenated and encoded by a location-specific dense layer with 512 output channels. In parallel, the shared geophysical features $\mathbf{a} \in \mathbb{R}^{8192}$ are passed through a second, location-specific dense layer with 512 output channels whose role is to extract the atmospheric information relevant to that particular gauge. The two 512-dimensional vectors are concatenated and processed by four dense layers with 1024 output channels, each followed by a SELU activation, dropout, and a residual connection, yielding the location-specific feature vector $\mathbf{x}_i \in \mathbb{R}^{1024}$. The residual (skip-connection) form is chosen for its stable gradient propagation (He et al., 2016); the SELU is retained from HIDRA3 for its self-normalising behaviour, and dropout curbs overfitting. Adding SLA as extra input columns of the first dense layer (rather than as a separate late branch) lets the low-frequency ocean signal propagate through the entire backbone. Those columns are initialised to zero, so the network starts numerically identical to the no-SLA model and learns to exploit SLA from the first epoch; this makes the with-SLA versus no-SLA comparison a clean, controlled ablation.

2.3.3 Feature fusion

The feature fusion module combines the location-specific features $\mathbf{x}_i$ into a single joint state vector $\mathbf{s} \in \mathbb{R}^{8192}$ that all tide gauges share. Its critical requirement is robustness to missing data, since the $\mathbf{x}_i$ are only defined for tide gauges with a valid input window. From each $\mathbf{x}_i$ two quantities are computed by location-specific dense layers: a partial state $\mathbf{s}_i \in \mathbb{R}^{8192}$ and a weight vector $\mathbf{w}_i \in \mathbb{R}^{8192}$. Each coordinate of $\mathbf{w}_i$ reflects how much tide gauge i contributes to that coordinate of the joint state; a gauge whose input is uninformative has its weight lowered, reducing its influence. The joint state is

$$\mathbf{s} = \sum_{i \in V} \hat{\mathbf{w}}_i \odot \mathbf{s}_i, \tag{1}$$

https://doi.org/10.5194/egusphere-2026-5073
Preprint. Discussion started: 31 August 2026

where $\odot$ is the element-wise product, V is the set of tide gauges with a valid input window this hour, and the coordinate-normalised weights are a per-coordinate softmax over the valid tide gauges,

$$\hat{w}_{i,j} = \frac{e^{w_{i,j}}}{\sum_{k \in V} e^{w_{k,j}}}. \tag{2}$$

Restricting the sum and the softmax to V ensures that tide gauges with missing input contribute nothing, so the joint state is always assembled from whatever observations are available. This is the mechanism by which a tide gauge with a data gap borrows information from its neighbours, and it is what lets HIDRA3 (and therefore MAROR) reconstruct any subset of tide gauges (including those with unobserved data at a given time) from those that are observed.

2.3.4 Sea-level regression

The regression head for tide gauge i receives the joint state $\mathbf{s}$ and the location-specific features $\mathbf{x}_i$ and produces the deterministic point estimate $\hat{\mu}_i$ of the sea level at the prediction lead time. Because some tide gauges may be under-represented in the joint state, or their $\mathbf{x}_i$ may be undefined (no valid input), a location-specific feature vector $\hat{\mathbf{x}}_i$ is formed: when gauge i has a valid input, $\hat{\mathbf{x}}_i$ is $\mathbf{x}_i$ passed through a dense layer with 1024 output channels; otherwise it is reconstructed from the joint state $\mathbf{s}$ by a separate dense layer with 1024 output channels. The vectors $\hat{\mathbf{x}}_i$ and $\mathbf{s}$ are concatenated and mapped by a final dense layer to $\hat{\mu}_i$. The one difference from HIDRA3 is the output width: HIDRA3 emits a 72 h forecast, whereas MAROR is a one-hour predictor ($\hat{\mu}_i \in \mathbb{R}$) that is applied autoregressively to reconstruct arbitrarily long series.

2.3.5 Generative ensemble

The last module turns the single reconstruction into a *distribution* of plausible reconstructions, from which the uncertainty is read. HIDRA3 does this with a Gaussian head: alongside the mean it predicts a standard deviation σ, so its uncertainty is always a symmetric bell curve. MAROR instead uses a *generative ensemble*: instead of describing the uncertainty with a prescribed distribution, it produces many alternative predictions (ensemble *members*), all equally plausible, and takes their scatter as the uncertainty. This lets the uncertainty be skewed or non-Gaussian (for example a long upper tail during a surge), which a single σ cannot represent.

A member is produced by injecting randomness. A noise vector $\mathbf{z} \in \mathbb{R}^{32}$ is drawn from a standard normal distribution (a fresh draw per member, shared by all tide gauges of that member) and mapped by a small multilayer perceptron to an embedding $\mathbf{e} \in \mathbb{R}^{128}$. For gauge i, a small *residual head* r_i then reads this noise together with the same features the mean uses, and outputs a perturbation that is added to the mean:

$$\hat{y}_i = \mathrm{sg}(\hat{\mu}_i) + r_i\big(\mathrm{sg}([\hat{\mathbf{x}}_i; \mathbf{s}]),\ \mathbf{e}\big). \tag{3}$$

If we put it in words, each member is constructed by adding the mean plus a noise-driven perturbation. Here $\hat{\mu}_i$ is the mean from the regression head, $[\hat{\mathbf{x}}_i; \mathbf{s}]$ are its features, and $\mathrm{sg}(\cdot)$ is the *stop-gradient* operator, which lets a value pass forward unchanged

https://doi.org/10.5194/egusphere-2026-5073
Preprint. Discussion started: 31 August 2026

but blocks the training signal from flowing back through it. A different $\mathbf{z}$ gives a different perturbation, hence a different member; drawing many of them (50 in our reconstructions) produces a cloud of values around the mean whose width is the predictive uncertainty. Applying the stop-gradient to both the mean and the features is the key design choice: the spread is learned only on top of these *detached* quantities, so training the ensemble can never move the mean or the shared backbone, and the central trajectory stays exactly the signal learned under the mean objective.

The spread head is trained so that this cloud has the right width at every lead time, and two ingredients make that work. First, the loss is the CRPS (Continuous Ranked Probability Score, Gneiting and Raftery (2007)), a standard scoring rule for probabilistic forecasts that measures how well the whole predicted distribution matches the observation: it rewards members close to the truth while penalising a spread that is too narrow or too wide. Because it is a *proper* scoring rule (minimised only when the forecast distribution equals the real one), minimising it teaches the ensemble the correct spread and not merely the correct centre. We use its *almost-fair* variant, which removes the bias that the plain CRPS estimator has when computed from a finite number of members (that bias would otherwise reward an artificially narrow ensemble). Second, the spread is trained not on a single one-hour step but on a *rollout*: the members are run autoregressively for 72 h, each feeding its own prediction back in as it advances, exactly as at reconstruction time. Training on the rollout is what teaches the ensemble that the uncertainty must grow with lead time, i.e. the band is narrow at the next hour and widens the further ahead the reconstruction runs. A per-station rescaling, applied afterwards during reconstruction, calibrates the final band.

2.4 Training

A separate network is trained for each user-defined region, always on the per-station standardised sea level so that the objective is region-independent: a station in a microtidal basin and one with a large tidal range contribute comparably to the loss regardless of their physical amplitude. Training proceeds in four phases, run in order and each initialised from the best state of the previous one (Sect. 2.4.5): the deterministic mean is first trained one hour ahead under teacher forcing (phase 1), then fine-tuned on its own self-fed rollout to remove a possible drift (phase 2); the mean is then frozen and only the stochastic spread head is trained on a self-fed ensemble rollout (phase 3); finally a per-station factor calibrates the uncertainty bands at reconstruction time (phase 4). All phases share the same optimisation settings: the AdamW optimiser with weight decay 10^{-3}, gradient-norm clipping at 1.0, and a cosine-annealed learning rate decaying to 10^{-2} of its initial value. Every phase early-stops on a validation metric and writes its best state to disk as it improves, so that an interrupted run never loses its best checkpoint and the exported weights are always the best state reached.

2.4.1 Phase 1: mean (teacher-forced MSE).

The one-hour predictor is trained to minimise the mean squared error between its estimate and the observed sea level one hour ahead, computed on the standardised target and masked wherever the observation is missing. The recent-history input is the *observed* sea level (teacher forcing), so this phase learns the instantaneous forcing-to-sea-level mapping without yet being exposed to its own errors. It runs at a learning rate of 10^{-4} for up to 200 epochs and is early-stopped (patience 15 epochs) on the validation RMSE, evaluated in meters after de-normalisation.

https://doi.org/10.5194/egusphere-2026-5073
Preprint. Discussion started: 31 August 2026

2.4.2 Phase 2: rollout with scheduled sampling (drift removal).

A network trained only under teacher forcing meets, at reconstruction time, an input distribution it never saw in training: its own imperfect past predictions. This mismatch, known as exposure bias, makes an autoregressive rollout accumulate a slow drift over the long reconstruction. Phase 2 tackles this issue by fine-tuning the mean on short self-fed rollouts that mimic the reconstruction loop. Each training example is a block of $K = 72$ consecutive hours: the sea-level history is seeded with the 72 h of *observed* sea level immediately preceding the block, and the network then advances one hour at a time through the block. At every step it predicts the next hour from the current 72 h history window (together with the tide, the atmospheric forcing and, where available, the SLA), the squared error of that prediction against the observed sea level is accumulated into the loss, and the history window is slid forward by one hour: the oldest hour is dropped and a new most-recent hour is appended before the next step.

Which value is appended at each step is the essence of scheduled sampling (Bengio et al., 2015). Independently for each gauge and each step, the appended hour is the network's own prediction with probability p and the observation with probability $1 - p$ (and always the prediction wherever the observation is missing, so that data gaps are handled exactly as they will be in reconstruction). The feedback probability p is ramped linearly from 0 to 1 over the first 8 epochs: the network starts in pure teacher forcing ($p = 0$, every appended hour is an observation) and is weaned gradually onto its own trajectory until, from the ninth epoch onward, it runs fully self-fed ($p = 1$), which is exactly the regime of the reconstruction. The fed-back prediction is detached from the computation graph, so the training signal is not back-propagated through the rollout; each step's gradient flows only through that step's own forward pass. The aim of the phase is thus to expose the mean to its own error distribution and let it learn to correct itself step by step, rather than to back-propagate through time, which also keeps the memory cost of a 72 h rollout at roughly that of a single forward pass. Only the mean and backbone weights are updated (the stochastic spread head is left untouched), at a smaller learning rate (2×10^{-5}) for up to 50 epochs (patience 8), early-stopped on the RMSE of a fully self-fed ($p = 1$) rollout over the validation sequences. The mean bias, the diagnostic most sensitive to drift, is driven close to zero by this phase.

2.4.3 Phase 3: generative ensemble (almost-fair CRPS).

With the drift-corrected mean frozen, the third phase trains only the noise embedding and the per-gauge residual head that turn latent noise into ensemble members (Sect. 2.3.5). The loss is the CRPS (Gneiting and Raftery, 2007) estimated over $M = 4$ members drawn per step: it combines a term rewarding members close to the observation with an inter-member term rewarding spread, so the ensemble cannot collapse onto the mean. Computed from a finite number of members, the plain CRPS estimator systematically rewards an artificially narrow ensemble; the *fair* CRPS removes this finite-ensemble bias (Ferro, 2014). We use the *almost-fair* variant (Lang et al., 2026) (with $\alpha = 0.95$), which interpolates between the plain and the fair estimators, keeping the ensemble spread nearly unbiased while retaining a usable training gradient. As in phase 2 the members are run over a $K = 72$ h self-fed rollout, with the feedback probability ramped over 10 epochs; training on the rollout is what teaches the spread to widen with lead time. Crucially, each of the M members carries its own latent noise and, during the rollout, feeds

https://doi.org/10.5194/egusphere-2026-5073
Preprint. Discussion started: 31 August 2026

back its own prediction, so the members diverge from one another as the rollout advances; this growing member-to-member spread is exactly the lead-time-dependent uncertainty the phase learns to calibrate. This phase runs at a learning rate of 10^{-4} for up to 15 epochs (patience 12), early-stopped on the validation rollout CRPS and monitored with the spread-skill ratio (which approaches 1 when the predicted spread matches the actual error) and the empirical coverage of the member quantiles.

2.4.4 Phase 4: post-hoc inflation.

A single multiplicative factor per station, fitted on the held-out validation years, rescales the distance of the ensemble bands from the mean so that a target fraction (90 %) of the validation observations fall inside them, restoring calibration while preserving the band asymmetry. It is applied at reconstruction time and never alters the mean; it is described together with the reconstruction procedure in Sect. 2.5.

2.4.5 Initialization

Phase 1 starts from randomly initialised weights, with one deliberate exception that keeps the central comparison of the paper clean. The columns of the first per-gauge dense layer that receive the SLA input are initialised to zero (Sect. 2.3.2), so at the first optimisation step the with-SLA network produces exactly the same output as the no-SLA network and departs from it only as it learns to exploit the altimetry; the full-versus-no_sla ablation therefore differs solely in whether SLA is available, not in the initial conditions. The subsequent phases are chained: phase 2 resumes from the best teacher-forced mean of phase 1, and phase 3 resumes from the drift-corrected weights of phase 2 with the mean and backbone frozen, so that the spread head is learned on top of a fixed central trajectory. All runs use a fixed random seed for reproducibility.

2.5 Reconstruction

Once trained, the one-hour predictor is turned into a reconstruction engine by applying it autoregressively over the target record. The reconstruction is run *fully self-fed*: no observed sea level is ever fed back into the network, so the trajectory is driven only by the astronomical tide and the atmospheric (and, when available, altimetric) forcing. This is what lets MAROR run continuously over the entire ERA5 period, from 1940 to the present, at every tide gauge and even in years where observations are sparse or absent; the train/validation/test split governs only which years are *scored*, not which are reconstructed, so the whole record is produced in a single unbroken rollout.

2.5.1 Cold start.

The rollout needs an initial 72 h of sea-level history which, at the start of the record, cannot come from observations. Each station's seed window is instead built from quantities available everywhere: the astronomical tide plus the inverse-barometer response to atmospheric pressure,

$$\eta_{\mathrm{ib}} = -\frac{P_{\mathrm{bar}} - P_{\mathrm{ref}}}{\rho g}, \tag{4}$$

https://doi.org/10.5194/egusphere-2026-5073
Preprint. Discussion started: 31 August 2026

where P_{bar} is the ERA5 mean sea-level pressure averaged over the grid cells within 150 km of the gauge, P_{ref} is the record-mean pressure, $\rho = 1025\,\mathrm{kg\,m^{-3}}$ and $g = 9.81\,\mathrm{m\,s^{-2}}$ (a sensitivity of about 1 cm per hPa). The seed is therefore the static tide-plus-inverse-barometer level, and from there the network takes over.

2.5.2 Autoregressive rollout.

An ensemble of $E = 50$ members is started from this common seed. At each hour every member is fed its current 72 h history window together with the ± 72 h windows of tide and forcing (and the SLA window where available) and a fresh draw of latent noise, and returns its estimate of the next hour; that value is appended to the member's own history, the window slides forward by one hour, and the step repeats. The forcing streamed in is always the real reanalysis; only the past-sea-level slot is self-fed, and each member feeds back *its own* prediction, so the members diverge as the rollout advances and their spread grows with lead time. At every hour the reconstruction stores, per station, the ensemble mean (the point reconstruction), the ensemble standard deviation, and the empirical member quantiles at the levels $\{0.05, 0.25, 0.5, 0.75, 0.95, 0.99, 0.999, 0.9999\}$, from which the uncertainty bands and a high-quantile ($q_{0.95}$) peak estimator for extremes are read.

2.5.3 Calibration and outputs.

The raw ensemble bands are finally calibrated by the post-hoc per-station inflation of phase 4 (Sect. 2.4): a single factor β per tide gauge, fitted on the held-out validation years, rescales the distance of the bands from the mean so that the nominal fraction of validation observations falls inside them, leaving the mean untouched. The product saved for each station is thus the gap-free hourly reconstruction (ensemble mean), its calibrated uncertainty bands and standard deviation, and the peak estimator, all in meters and spanning the full ERA5 record. Skill is assessed only on the held-out test years, against two baselines that isolate what the network adds beyond the trivial forcing response: the static tide-plus-inverse-barometer level for the total sea level, and a zero-surge (tide-only) prediction for the de-tided residual. Several skill metrics are reported with the results: RMSE, bias and correlation, the empirical coverage of the bands, the spread-skill ratio, and extreme-event scores (probability of detection, false-alarm ratio, and the reproduction of event maxima) on the de-tided surge.

3 Data and study regions

MAROR is a region-independent engine: nothing in Sect. 2 is tied to a particular dataset or geographical domain. This section describes the specific observations and forcing used to show it, and the set of regions over which it is trained and evaluated.

3.1 Tide gauge records and quality control

The target sea level is taken from version 3 of the GESLA database (Global Extreme Sea Level Analysis; Haigh et al. 2023), a global compilation of high-frequency sea-level records contributed by national and regional authorities. We use its hourly (or higher-frequency) records. Tide gauges are selected to represent open-coast conditions (thus discarding stations that are flagged as riverine or lacustrine) and to provide enough data to train on, requiring at least 25 years of valid hourly observations.

https://doi.org/10.5194/egusphere-2026-5073
Preprint. Discussion started: 31 August 2026

The resulting tide gauges and their distribution across regions are summarised in the study-regions subsection below and listed station by station in the supplementary material.

Raw tide gauge records carry data gaps, isolated spikes, datum shifts and instrumental drift, all of which must be removed before the series can be used for training purposes. In other words, only those periods of data with good quality are usable for training.

The procedure runs as follows. Each record is first converted into hourly sampling by averaging any sub-hourly values within each hour, and segments sampled more coarsely than hourly are discarded, since they cannot inform an hourly model and interpolating them would fabricate data. A conservative outlier detector is then applied to the total sea level (at six standard deviations) to safeguard the datum estimation and the tidal oscillations in macrotidal regimes; datum shifts are corrected by change-point detection (the PELT algorithm; Killick et al. 2012), and any residual linear trend is removed by robust (bisquare) regression. A single harmonic analysis is then fitted to the real observations with the UTide package (Codiga, 2011), which selects the tidal constituents automatically from the length of each record (by the Rayleigh criterion), applies nodal and satellite corrections, includes a secular trend term, and solves for the harmonic constants by ordinary least squares. This harmonic tide, predicted from the fitted constants, is the astronomical tide used as a model input (Sect. 2.2) throughout this work. The surge is then formed as the observed level minus this tide, only at hours where an observation exists; a second outlier check, now more effective because the tidal variance is gone, is run on the surge (at five standard deviations), short surge gaps (under 5 h) are filled by linear interpolation, and valid segments shorter than one month are discarded. Finally the cleaned total sea level is recomposed as the surge plus the astronomical tide, so that the short interpolated gaps follow the tidal motion. The astronomical tide is defined at every hour, but the observed sea level and surge remain undefined wherever no observation exists. These unobserved periods are excluded from the training, validation and test sets, so the network is never trained or scored on a tide-only value (the tide-plus-inverse-barometer fill of Sect. 2.5 is used only to cold-start the reconstruction and is likewise never scored).

3.2 Atmospheric and oceanic forcing (ERA5)

The atmospheric and oceanic forcing is taken from the ERA5 reanalysis (Hersbach et al., 2020) at hourly resolution, spanning 1940 to the present. The native $0.25°$ fields are regridded to the $1°$ grid on which the network operates (Sect. 2.2), from which we retain the eight channels described there. The mean wave direction, being a circular quantity, is split into its sine and cosine so that the network sees a continuous two-channel representation rather than a discontinuity at $0°/360°$. Each channel is then standardised over the domain; land cells are set to the standardised mean (zero), a neutral fill that the geophysical encoder learns to ignore.

Each region is forced by its own ERA5 window, a rectangular box drawn by hand for that region (listed in the study-regions subsection below). The boxes are sized by three considerations: a sufficient spatial buffer around the region's tide gauges; an extension *upstream*, in the direction from which the region's storms typically arrive, so that the encoder sees the forcing before it reaches the coast; and the avoidance of large land areas, which carry no useful signal and would only enlarge the grid.

https://doi.org/10.5194/egusphere-2026-5073
Preprint. Discussion started: 31 August 2026

A caveat accompanies the earliest part of the record: ERA5 is of lower quality before 1979, prior to the assimilation of satellite observations, so the first decades of the reconstruction rest on a sparser observational basis and should be interpreted with correspondingly lower confidence.

3.3 Satellite altimetry

The optional SLA input (Sect. 2.2) is derived from the SSALTO/DUACS Delayed-Time Level-4 altimetry product distributed by the Copernicus Marine Service (E.U. Copernicus Marine Service Information (CMEMS), 2024), produced with the DUACS processing system (Taburet et al., 2019): a global daily map of sea-level anomaly at $1/8°$ resolution, built by merging all available altimeter missions. From it we construct one daily SLA series per tide gauge by averaging the gridded anomaly over the ocean cells lying within a great-circle radius of 100 km of the tide gauge; for the few tide gauges that would otherwise enclose fewer than five ocean cells (for instance in narrow embayments), the radius is grown until it does, up to a cap of 300 km. Averaging over a neighbourhood rather than sampling a single cell suppresses altimetry noise, avoids less representative grid points, and yields a robust local estimate of the low-frequency ocean sea level. In addition, its linear trend is removed so that the SLA time series represents the ocean variability rather than the long-term rise in mean sea level. The network reads this series as a window of 30 samples at daily spacing ending at the prediction hour (Sect. 2.2); the daily product is interpolated in time onto the hourly time reference of the tide gauges, which, because the SLA varies slowly, introduces no meaningful error.

Two properties of the product matter for its use here. First, the DUACS anomaly is expressed relative to a fixed 1993 to 2012 mean, so the series contains the ocean mesoscale and interannual variability (boundary currents, eddies, basin-scale adjustment) that the atmospheric reanalysis cannot contain, which is precisely the signal MAROR is meant to gain from it. Second, the product already includes the dynamic atmospheric correction, so it does not carry the high-frequency, inverse-barometer pressure response that the network already receives through the ERA5 mean sea-level pressure; the two inputs therefore do not double-count the atmospheric signal.

Since altimetry is only available from 1993 onward, before that period the SLA input is set to zero (the standardised mean, a neutral fill), consistent with the zero-initialised SLA columns described in Sect. 2.3.2: the network falls back smoothly to its no-SLA behaviour in the pre-altimetry era, and the with-SLA versus no-SLA comparison remains a controlled ablation over the satellite period.

3.4 Study regions and temporal splits

A separate network is trained for 11 different regions (Fig. 2). These are chosen to encompass a wide range of coastal settings and, at the same time, to keep each domain small enough for the shared geophysical encoder to compress. The eleven regions cover regimes that range from extratropical storm surge on wide continental shelves (North Sea, Baltic Sea, West British Isles, Spain/France, US East Coast) through semi-enclosed seas of the Western Mediterranean and Japan Sea, to settings dominated by tropical cyclones and strong ocean currents (Gulf of Mexico, with its hurricanes and Loop Current; Japan, with its typhoons and the Kuroshio) and by eastern-boundary and Southern-Ocean variability (US West Coast, Western Australia, Southern

https://doi.org/10.5194/egusphere-2026-5073
Preprint. Discussion started: 31 August 2026

Table 1. The eleven study regions: number of tide gauges N, ERA5 forcing box ($[\mathrm{lon}_{\min}, \mathrm{lon}_{\max}]$, $[\mathrm{lat}_{\min}, \mathrm{lat}_{\max}]$ in degrees, with negative values west and south), record period, and the number of years in the training/validation/test split. Japan is additionally trained as northern (57) and southern (42) sub-networks split at 34.7°N.

Region	N	ERA5 box (lon, lat)	Period	Train/val/test
North Sea	60	$[-8, 14]$, $[46, 65]$	1940–2020	57/12/12
Baltic Sea	37	$[6, 31]$, $[52, 67]$	1940–2020	57/12/12
West British Isles	24	$[-17, 3]$, $[45, 62]$	1958–2020	45/9/9
Spain/France	13	$[-16, 2]$, $[34, 51]$	1961–2020	41/9/9
Mediterranean	10	$[-4, 17]$, $[34, 46]$	1960–2020	34/7/7
US East Coast	38	$[-78, -58]$, $[31, 47]$	1940–2020	57/12/12
Gulf of Mexico	23	$[-98, -76]$, $[15, 32.5]$	1944–2020	54/11/11
US West Coast	27	$[-135, -116]$, $[30, 52]$	1940–2020	57/12/12
Western Australia	11	$[103, 123]$, $[-40, -22]$	1966–2019	38/8/8
Southern Australia	29	$[127, 159]$, $[-49, -30]$	1963–2019	41/8/8
Japan	101	$[124, 154]$, $[25, 48]$	1940–2019	52/11/11
Total	373			

Australia). In total the eleven regions comprise 373 distinct tide gauges (their numbers, forcing boxes and record periods are summarised in Table 1 and listed station by station in the supplementary material).

Japan is additionally trained as two latitudinal sub-networks, split at 34.7°N into a northern and a southern subset (Fig. 2c), because the sea level south of that boundary is strongly influenced by the presence of the Kuroshio current; together with the eleven regional networks this gives thirteen trained configurations. Each region is forced by its own ERA5 box (Sect. 3.2), whose extent is given in Table 1 and is shown in Fig. 2; the northern and southern Japan sub-networks share the full Japan box.

Within each region the record is partitioned by whole years into training, validation and test sets, balanced by extremes so that each set includes a comparable share of extreme events. Each year is scored by its annual maximum surge, taken across all the region's tide gauges, so that the severity reflects the largest event seen anywhere in the region that year; the years are then ranked by this severity and interleaved across the three sets in an approximate 70/15/15 proportion, with the single most severe year assigned to training so that the network sees the largest event. Because the balancing is performed independently per region, the split years differ from region to region (Table 1 gives the sizes; the exact years are listed in each region's configuration). Note that the split defines only which years are used to train and to score the model, but the reconstruction always spans the full ERA5 record (Sect. 2.5).

3.5 Computational cost

All networks were trained on the TalaIA cluster (University of the Balearic Islands), each region on a single NVIDIA GPU. Because MAROR uses a separate set of dense layers to every tide gauge, its parameter count grows linearly with the number of tide gauges in the region, at about 34.8 million parameters per tide gauge. The smallest configuration (Mediterranean, 10 tide gauges) therefore holds 0.36 billion parameters and the largest (Japan, 101 tide gauges) 3.5 billion. Memory is consequently the binding constraint and it scales with the number of tide gauges and not with the length of the record. Regions were assigned

https://doi.org/10.5194/egusphere-2026-5073
Preprint. Discussion started: 31 August 2026

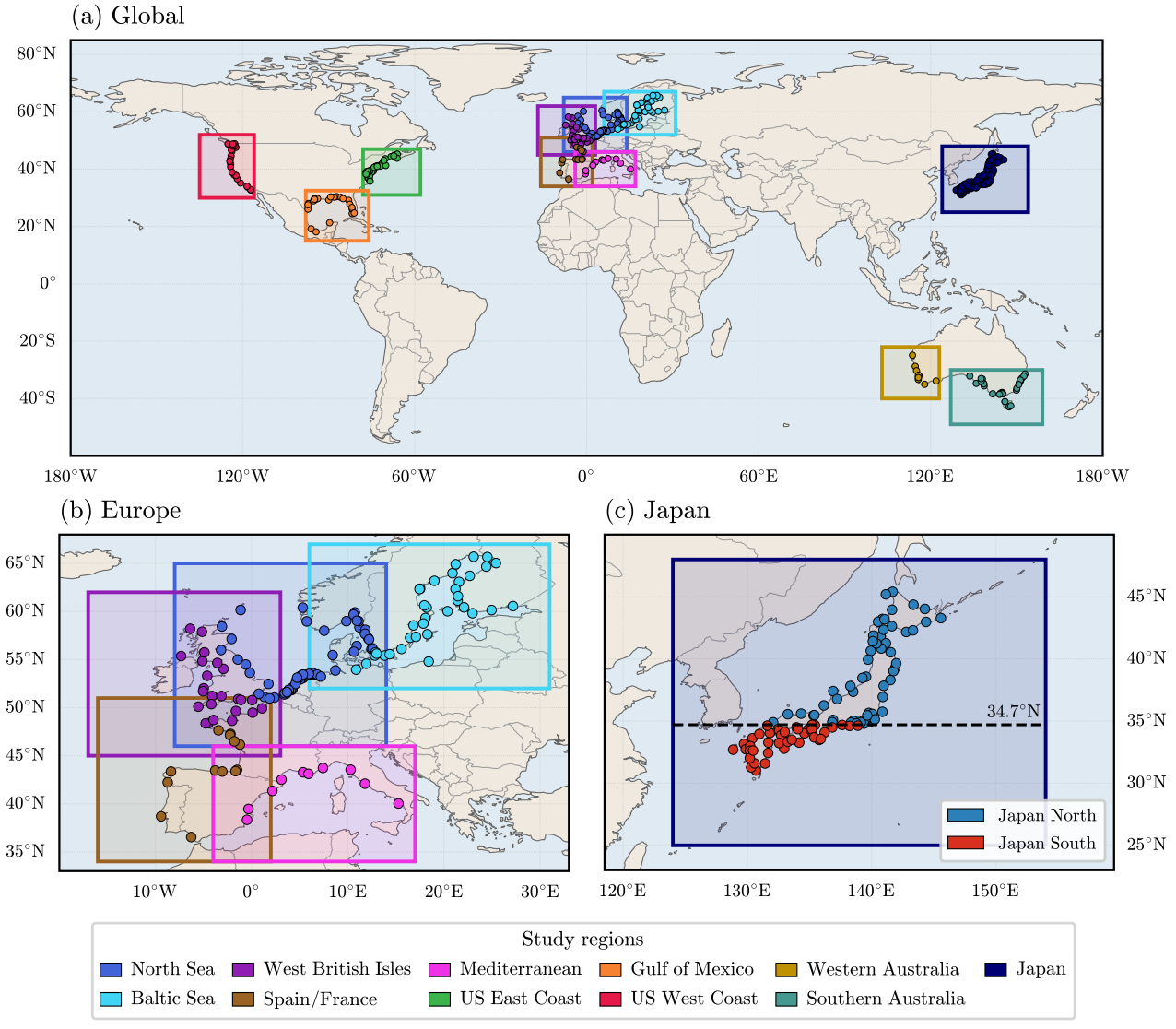

Figure 2. The eleven study regions. (a) Global distribution; (b) European regions; (c) Japan, with the 34.7°N boundary between the northern and southern sub-networks. Coloured rectangles are the ERA5 forcing boxes and coloured dots are the tide gauges, one colour per region.

https://doi.org/10.5194/egusphere-2026-5073
Preprint. Discussion started: 31 August 2026

to hardware accordingly: those with up to 29 tide gauges ran on an L40S (48 GB), those with 37 to 60 on an H200 (141 GB), and the full Japan network on a B200 (180 GB). A single GPU was sufficient in every case, and no model or data parallelism was required.

Training a region through the four phases took between 1.6 h (Mediterranean, 10 tide gauges) and 66.9 h (Japan, 101 tide gauges), the thirteen configurations totalling about 247 GPU-hours. Nearly all of this is spent in the first, teacher-forced phase, which accounts for roughly 90 % of the training time in the larger regions; the rollout, ensemble and calibration phases together add only a few hours.

Reconstruction is markedly cheaper than training. Producing the full record for one region, that is 756 768 hourly steps per tide gauge from 1940 to 2026 with an ensemble of $E = 50$ members, took between 1.2 h (Mediterranean) and 7.1 h (Japan), about 51 GPU-hours for all thirteen configurations. Although the rollout is strictly sequential in time, the tide gauges and the ensemble members advance in parallel within each hour, so the cost is set by the length of the record and only weakly by the number of tide gauges. Once trained, a network therefore reconstructs 86 years of hourly sea level at every tide gauge of its region in a few hours on one GPU, and can be re-run at negligible cost whenever the forcing is extended or revised. The whole study, thirteen trained configurations and their reconstructions, amounts to approximately 299 GPU-hours, around 12.5 days.

4 Results

This section reports the overall performance of reconstructed sea levels across all 373 tide gauges. It identifies the regions where MAROR reconstructs the surge best, and examines where it has a poorer performance and why, separating the cases in which the low skill reflects a genuine physical limit, such as the presence of unseen tropical cyclones or the slow ocean variability that the atmospheric forcing cannot resolve, from those in which it reflects low-quality observations rather than a failure of the model. In the latter, we show how MAROR can be used as an efficient and reliable quality control procedure for tide gauge observations.

Beyond the figures shown below, every tide gauge can be examined individually in an interactive companion to this paper at https://maror.angelamores.org, where the reconstructed and observed storm surge are compared over the full 1940 to 2026 record together with the per-gauge skill metrics, Taylor diagram and quantile–quantile plot, so that the reconstruction can be reviewed tide gauge by tide gauge.

Unless otherwise stated, all scores are computed on the storm surge over each region's test years. Scoring the surge rather than the full sea level avoids that the tidal signal, which is highly predictable, inflates the scores artificially.

4.1 Reconstruction skill across regions

Across the 373 tide gauges MAROR reconstructs the hourly surge with a median correlation of 0.88 and a median skill score of 0.52, where the skill score $1 - \mathrm{RMSE}/\sigma_{\mathrm{obs}}$ (σ_{obs} the standard deviation of the observed surge) equals one for a perfect reconstruction and zero for one no better than predicting the time-mean. The reconstruction is essentially unbiased: the median bias stays within ± 0.011 m in all regions, which confirms that the scheduled-sampling rollout of the second training phase

https://doi.org/10.5194/egusphere-2026-5073
Preprint. Discussion started: 31 August 2026

(Sect. 2.4) removes the slow drift that a purely teacher-forced network would otherwise accumulate over a multi-decadal, fully self-fed reconstruction. Figure 3 maps these results: the per-tide-gauge correlation (top row) and skill score (middle row) are shown as coloured dots for each tide gauge, and a Taylor diagram (bottom row) summarises, for the same tide gauges, the joint behaviour of correlation and normalised standard deviation.

The skill is highest in the extratropical storm surge regimes of wide, shallow continental shelves, where the surge is an almost direct response to wind and pressure that ERA5 reanalysis resolves well. This is the case of the North and the Baltic Seas that reach median correlations of 0.94 and 0.93 (skill 0.66 and 0.63), respectively, and the US East Coast with 0.92. Conversely, it is lowest where a large part of the storm surge variance comes from slow ocean dynamics that the atmospheric forcing does not contain. This occurs in southern Japan, where sea level variability is highly influenced by the Kuroshio and it is furthermore struck by typhoons; this is the single hardest region (median correlation 0.71, skill 0.29). The Japan column of Fig. 3 evidences this spatial pattern, with a tide-gauge colour shift across the $34.7°$N Kuroshio front from the better-reconstructed north (median correlation 0.82) to the harder south (0.71), which is precisely the boundary along which the two Japan sub-networks are trained.

The Taylor diagrams expose a second, consistent feature. The normalised standard deviation lies below one almost everywhere (median 0.92), i.e. the reconstruction tends to slightly under-represent the observed variability. This shortfall is small in the atmospherically driven regions (median $\sigma_{\mathrm{mod}}/\sigma_{\mathrm{obs}} = 0.93$) but more pronounced in the ocean-driven ones (median 0.89, falling to 0.83 in southern Japan). A plausible reading is physical rather than a training artefact: part of the surge variance in these regions arises from slow ocean adjustment, such as boundary currents and mesoscale eddies, that the atmospheric forcing does not resolve. The SLA input is designed to supply precisely this low-frequency ocean signal, yet the reconstructed variability still falls short, which suggests that the altimetry, as ingested here, is not sufficient to fully recover this part of the variance.

Japan offers a related, and perhaps more instructive, view of the same limitation. We trained it both as a single regional network and as two latitudinal sub-networks split at the $34.7°$N Kuroshio front (Sect. 3.4), to test whether separating the two regimes helped. Splitting improves the reconstruction over the same tide gauges, with the median correlation rising from 0.69 to 0.75 and the skill score from 0.27 to 0.33, even though each sub-network then draws on fewer tide gauges. This indicates that the sharp dynamical contrast across the Kuroshio front outweighs the benefit of pooling more stations: asking a single shared representation to span two markedly different regimes degrades it, whereas separating them lets each network specialize. The information that a tide gauge borrows from its neighbours through the fusion mechanism (Sect. 2.3.3) is therefore most useful within a physically coherent regime. Accordingly, we keep Japan split, with the northern sub-network (median correlation 0.82) clearly outperforming the southern one (0.71).

4.2 Reproducing surge variability and its extremes

Having established the overall skill, we now zoom to individual tide gauges to show examples of what a successful reconstruction looks like across scales, from the hourly variability to the extreme events. Figure 4 collects four tide gauges from four different regions and surge regimes: The Battery on the US East Coast, Esbjerg in the macrotidal North Sea, A Coruña

https://doi.org/10.5194/egusphere-2026-5073
Preprint. Discussion started: 31 August 2026

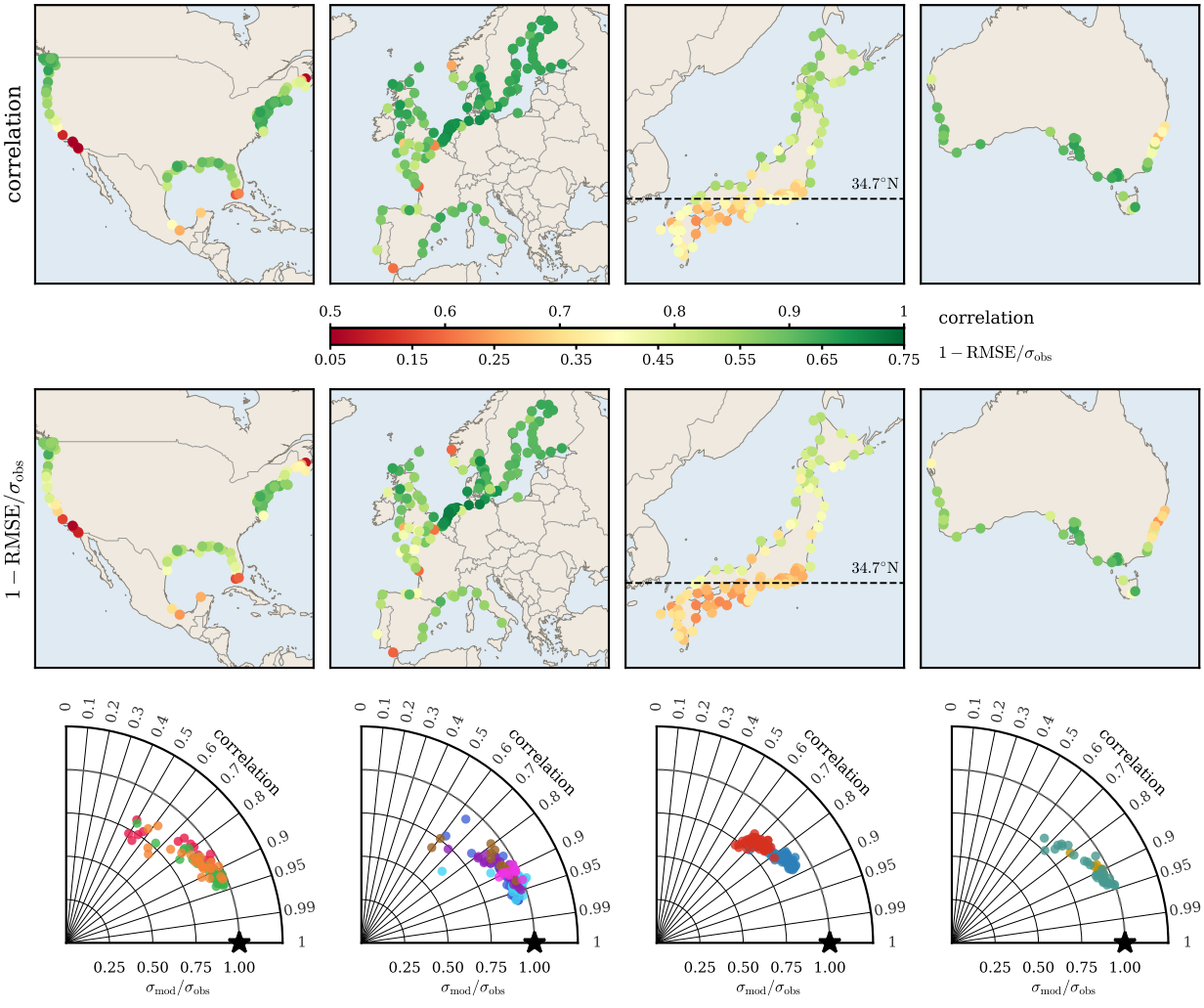

Figure 3. Per-tide-gauge reconstruction skill of MAROR on the storm surge over the test years, grouped into four geographic columns (North America, Europe, Japan, Australia). Top row: correlation. Middle row: skill score $1 - \mathrm{RMSE}/\sigma_{\mathrm{obs}}$. Each tide gauge is a dot whose fill encodes the metric on the shared red-to-green colour bar (correlation 0.5 to 1 on top; the skill score over its data range on the bottom). The Japan maps mark the 34.7°N boundary (dashed) between the northern and southern sub-networks, as in Fig. 2. Bottom row: Taylor diagram for each column, with the angle giving the correlation, the radius the normalised standard deviation $\sigma_{\mathrm{mod}}/\sigma_{\mathrm{obs}}$, and the black star the observations (correlation 1, unit standard deviation); each point is coloured by its region in the palette of Fig. 2.

https://doi.org/10.5194/egusphere-2026-5073
Preprint. Discussion started: 31 August 2026

on the Atlantic coast of the Iberian Peninsula, and Wallaroo in Southern Australia. All four combine good data quality with a surge that the atmospheric forcing resolves, and MAROR reconstructs them with correlations of 0.92, 0.95, 0.90 and 0.94 and skill scores of 0.60, 0.69, 0.55 and 0.65, respectively. Consistently with this, their normalised standard deviation stays close to one (0.94 to 0.96). Reaching this performance does not require a long record: The Battery and Esbjerg draw on 77 and 68 years of observations, whereas A Coruña and Wallaroo are reconstructed equally well from only 28 and 36 years, so a long observational record is not a prerequisite for an accurate reconstruction.

The quantile–quantile plots (left column) show the reconstructed and observed hourly surge quantiles falling along the identity line over essentially the whole range, so the reconstruction reproduces the full distribution of the surge and not merely its central spread. The quantiles computed over every observed hour (blue) and those restricted to the test years (red) lie on top of each other, which indicates that the fit does not degrade on the years the network never saw during training. Importantly, the agreement extends into the upper quantiles, where the surge is largest, confirming that the model follows the tail of the distribution and not only its bulk.

The centre column shows the surge return levels. The generalized extreme value distribution fitted to the reconstructed annual maxima over the observational period (green) matches the one fitted to the observations (red) within the bootstrap confidence bands at all four tide gauges, so MAROR recovers the extreme-value statistics and not just the individual peaks. The added value of the reconstruction appears in the blue curve, fitted to the full 1940 to present record: by using eighty-seven years of annual maxima rather than the few decades that carry observations, it narrows the confidence bands and stabilises the long return periods that a short record estimates poorly. Because this curve samples years without observations, some of its plotting positions exceed the largest surge present in the observed sample. A Coruña provides an illustration of this fact. Its largest reconstructed annual maximum, the uppermost blue point in the return-level panel, is a surge of 0.81 m on 15 February 1941, a year for which the record does not have any observation available. The event was therefore never available to the network as a training target, yet MAROR reconstructs it from the atmospheric forcing alone (the pre-1993 period carries no altimetry either), and its amplitude exceeds the largest surge in the observed sample. In the reconstruction the surge builds over about a day and peaks in the evening of 15 February, and the same storm appears about one hour later and considerably larger at nearby Vigo, where it reconstructs to 1.09 m, likewise the highest value of that record and likewise unobserved (see Fig. S1). This mid-February 1941 Atlantic storm is also reproduced by the independent hydrodynamic hindcast of Agulles et al. (2024) (their Fig. A18), which reaches about 0.65 m at A Coruña. The nature of the storm behind it is more striking still. As documented in Fig. A18 of Agulles et al. (2024), the 1941 surge was driven by a rare storm of hurricane-like character that reached these coasts, an event with no analogue in the regional record. That MAROR reconstructs it accurately, having never been exposed to a comparable forcing pattern during training, is a clear demonstration of its ability to generalise well beyond the conditions it learned from.

Finally, the event windows (right column) confirm that the model reproduces individual extreme events, both the largest surge of the entire record (top, which falls in a training year) and the largest among the test years (bottom). In each case the reconstruction follows the shape and the timing of the peak, and the 5 to 95 % ensemble band brackets the observed surge throughout the event. The reconstructed event maximum reaches between 88 and 104 % of the observed one across the four

https://doi.org/10.5194/egusphere-2026-5073
Preprint. Discussion started: 31 August 2026

tide gauges, so the peak is captured to within a few centimetres; this small, well-behaved misfit is the consequence of the forcing-limited underestimation examined next. Taken together, these results show that, where the data are of good quality and the forcing resolves the surge, MAROR is accurate from the distribution as a whole down to the individual extreme event.

4.3 Physical limits set by the forcing

In this section we examine how well MAROR reconstructs the parts of the sea-level record that are not well represented in the atmospheric forcing, through two contrasting examples, the peaks driven by tropical cyclones and the low-frequency variability of the ocean.

We begin with the ability of MAROR to reproduce the peaks driven by tropical cyclones. Figure 5 shows every tropical cyclone affecting a tide gauge across the regions that experience them, keeping the data for which the storm reached Saffir–Simpson category 1 or above within 150 km of the tide gauge, which yields 1903 events identified from the IBTrACS best track archive. Across this set MAROR reconstructs the surge peaks with a correlation of 0.91 and a median of 89 % of the observed peak, so even these short lived and spatially compact extremes are captured to within about a tenth of their amplitude, on average. It is worth noting that tropical cyclones drive only a minority of the strongest surges in these basins, about 13 % of the largest events in the Gulf of Mexico and 31 % in southern Japan, with extratropical storms accounting for the rest, yet tropical cyclones are the most extreme and most localised events.

The recovery of these peaks is represented by the ratio between reconstructed and observed values. The results (central panel of Fig. 5) display a median ratio staying between 0.86 and 0.90 from category 1 to categories 4 and 5 combined, so the nominal strength of a cyclone does not by itself predict how much of its peak is recovered. The deficit is instead regional (right panel). Along the US East Coast the peaks are recovered almost in full, with a median ratio of 0.96, whereas the Gulf of Mexico and northern Japan are the most underestimated at 0.84, and southern Japan lies in between at 0.91. This pattern points to local factors such as coastal geometry, bathymetry and the typical size of the storms that affect each coast, rather than to the category of the cyclone.

When the events are sorted by the distance to the cyclone (Fig. S2), the median peak ratio rises monotonically with this distance, from 0.80 for direct hits within 25 km to 0.93 beyond 100 km, and its spread narrows at the same time. This indicates that MAROR underestimates the surge more strongly the closer the cyclone passes to the tide gauge, whereas cyclones that pass farther away are, on average, better reproduced. The main reason lies in the forcing. For a direct strike the compact core of the cyclone and its steep pressure and wind gradients are underrepresented in the atmospheric reanalysis, so the forcing close to the eye is too weak, whereas the broader and weaker surges of distant passes are better represented by the forcing and, therefore, better reconstructed.

This limitation is attributed to the forcing rather than introduced by the network. The atmospheric reanalysis itself underestimates the intensity of tropical cyclones, so hydrodynamic surge models driven by it fall short of the observed peaks even at much finer resolution, and global surge products blend in parametric cyclone winds to recover the extremes (Dullaart et al., 2020; Muis et al., 2020). Refining the spatial resolution of the forcing on our side did not help either. A test at 0.5° did not

https://doi.org/10.5194/egusphere-2026-5073
Preprint. Discussion started: 31 August 2026

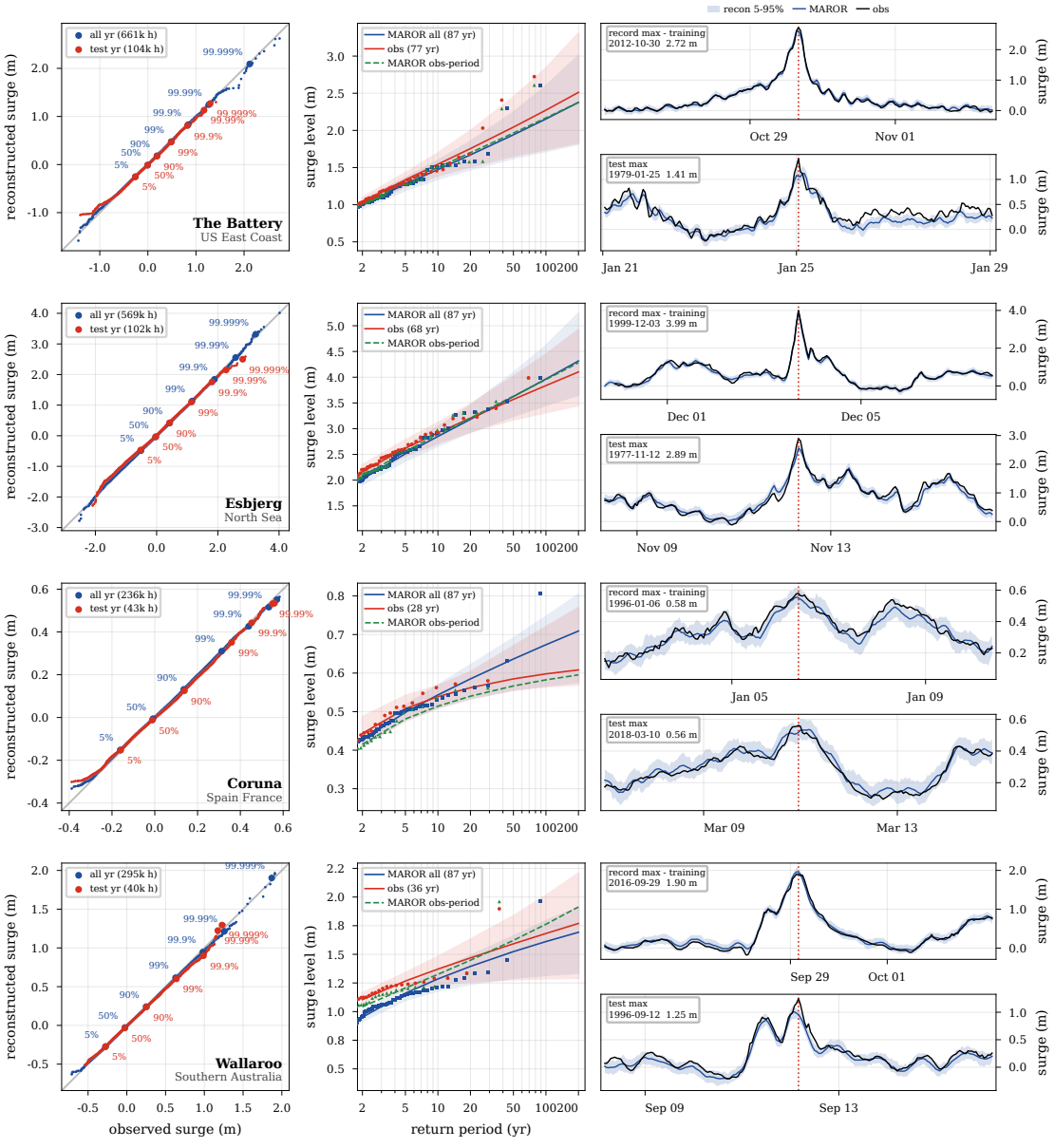

Figure 4. Reconstruction skill of MAROR at four representative tide gauges, one per row: The Battery (US East Coast), Esbjerg (North Sea), A Coruña (Spain–France) and Wallaroo (Southern Australia). Left column (quantile–quantile plots): reconstructed versus observed hourly surge quantiles. Blue points use *every hour with an available observation* (number of hours in the legend); red points use only the hours belonging to the test years. Centre column (surge return levels): generalized extreme value (GEV) distributions fitted by maximum likelihood to annual maxima of surge (years with at least 50 % hourly coverage), with 90 % bootstrap confidence bands and empirical (Weibull) plotting positions. Three estimates are shown: *obs* (red) from the observed annual maxima over the years with observations; *MAROR all* (blue) from the reconstructed annual maxima over the full reconstruction period 1940–present (87 years), including years without observations; and *MAROR obs-period* (green dashed) from the reconstructed annual maxima restricted to the timestamps that have observations (i.e. the same sample as *obs*). Because *MAROR all* spans the whole 1940–present period, its plotting positions can lie above the largest surge shown in the Q–Q panel, which is limited to observed hours. Right column (event windows, ±4 days): the top panel shows the largest surge of the entire record (*record max*, which falls in a training year, as labelled) and the bottom panel the largest surge among the test years. Black is the observed surge, blue the MAROR reconstruction and the light-blue band its 5–95 % quantile interval.

https://doi.org/10.5194/egusphere-2026-5073
Preprint. Discussion started: 31 August 2026

improve the reconstructed peaks over the 1° configuration. We therefore retained the 1° forcing, which greatly reduces the number of parameters and the training and reconstruction time without degrading the reconstructed extremes.

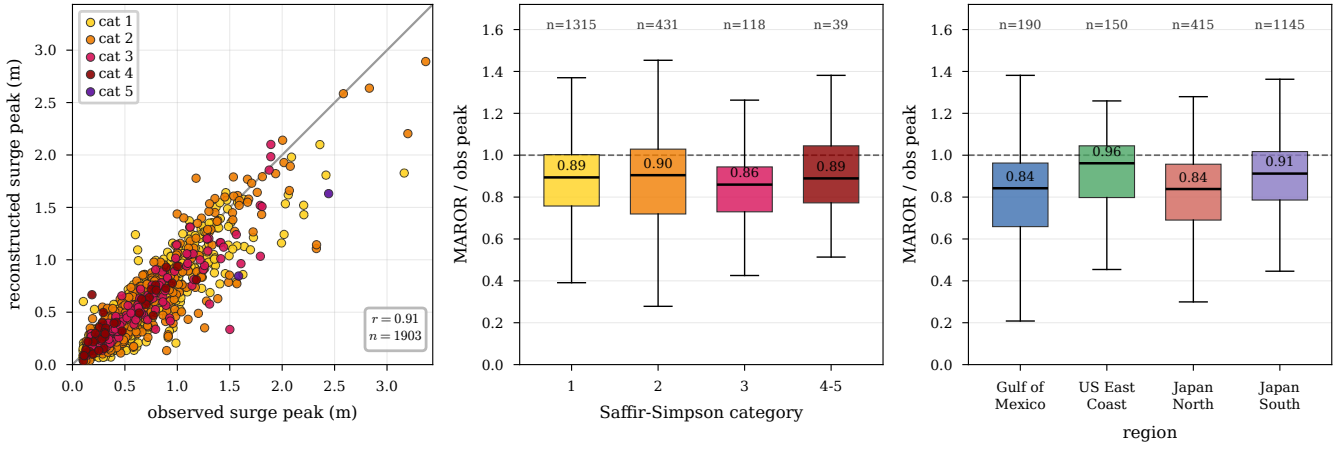

Figure 5. Reconstruction of tropical-cyclone storm-surge peaks. Each point represents the signal of a tropical cyclone on a tide gauge, retained when the storm reached Saffir–Simpson category 1 or above within 150 km of the tide gauge (1903 events across the Gulf of Mexico, the US East Coast, northern and southern Japan, and Western Australia; cyclone tracks and intensities from IBTrACS). For each event the storm-surge peak observed at the tide gauge is compared with the MAROR reconstruction, taken as its maximum within ±24 h of the observed peak. Left: reconstructed against observed storm-surge peak, coloured by Saffir–Simpson category, with the one to one line in grey and the Pearson correlation and sample size inset. Centre: the ratio of reconstructed to observed peak grouped by category (categories 4 and 5 combined). Right: the same ratio grouped by region merging all categories. In the box plots the central line is the median (annotated), the box spans the interquartile range and the whiskers extend to 1.5 times that range, with outliers omitted. Supplementary Fig. S2 breaks this ratio down by the distance of the cyclone at closest approach.

Another limitation lies in the low-frequency band, where part of the coastal sea-level variability may be set not by the local weather but by ocean dynamics such as steric expansion and western boundary currents that are independent of the atmospheric forcing. The satellite-altimetry input is intended to help with this signal, and we now examine how well MAROR reconstructs the low-frequency variability and how much the altimetry actually adds.

We isolate the low-frequency band by taking monthly means of the storm surge over the altimetry period (1993 to present) with the mean seasonal cycle removed, and we measure how much of its variance the reconstruction recovers through the ratio $\sigma_{\mathrm{mod}}/\sigma_{\mathrm{obs}}$. To quantify what the altimetry contributes, we compare the full network against a second version trained without the satellite-altimetry input, which we call the no-SLA configuration. At most tide gauges MAROR already reproduces this low-frequency variance well without altimetry, with a variance ratio close to 0.95. The reasons are because, on the one hand, the sea-surface temperature partially represents the slow steric contribution; on the other, low frequency pressure fields are largely responsible of this variability. Apart from the single exception discussed below, adding the altimetry shifts the median low-frequency variance ratio by at most 0.05 in any region, so it changes the picture very little.

The clear exception is southern Japan, the region already identified as the hardest in Sect. 4.1. There the low-frequency variability is dominated by the Kuroshio and its meanders, a mesoscale ocean signal that the atmospheric forcing does not

https://doi.org/10.5194/egusphere-2026-5073
Preprint. Discussion started: 31 August 2026

contain. Without altimetry MAROR recovers 86 % of the low-frequency variance and correlates with the observations at 0.77, whereas adding the altimetry raises the variance ratio to 0.98 and the correlation to 0.86. This is the one region where the altimetry makes a decisive difference, and it does so precisely where the missing signal is oceanic rather than atmospheric.

The low-frequency result therefore mirrors the tropical-cyclone one. Where the coastal sea level is driven by the atmosphere, MAROR reconstructs it well from the forcing alone, at both high and low frequency. Where it is instead driven by ocean dynamics that the forcing does not carry, the reconstruction falls short, and the satellite altimetry helps only to the extent that it samples that missing signal, most clearly in the Kuroshio region.

4.4 MAROR as a data-quality diagnostic

Because MAROR is trained jointly on many tide gauges and driven only by the atmospheric forcing, it also provides a way to control the quality of the records themselves. The reconstruction is, by construction, physically plausible and free of instrumental artifacts, so it can be read as an independent reference against which each observed record is compared. A tide gauge whose reconstruction skill is anomalously low, while its neighbours under the same forcing are reconstructed well, is unlikely to be suffering from a limitation of the forcing or of the network, and more likely to be carrying a defect in the observation. We illustrate the idea with Calais, which is the least well reconstructed of the sixty North Sea tide gauges, with a test correlation of 0.62, even though Dunkerque, only 40 km away, reaches 0.85 and Dover across the Strait reaches 0.92.

Comparing the observation with the reconstruction shows the problem. Figure 6a shows the storm surge at Calais over the full record: the observed and reconstructed surge agree from the mid-1990s onward, but between the late 1960s and the early 1980s the observation carries a systematic excess of variance that the reconstruction does not share. The nature of that excess becomes clear in the daily-scale window. In a calm week within the suspect period (Fig. 6c) the de-tided observation oscillates at tidal frequencies with an amplitude up to 1 m, while the reconstruction stays almost flat as the weather demands; in a calm week outside the suspect period (Fig. 6d) the two agree closely. A true storm surge is not likely to have spectral lines at the tidal frequencies, so a tide-coherent oscillation in the de-tided record seems to be the signature of a time-base error in the tide gauge clock: a small timing offset leaves, after the astronomical tide is removed, a residual proportional to the rate of change of the tide. This is a common issue at many tide gauge time series, particularly in historical records.

This signature can be turned into a quantitative detector. A clock offset δ makes the instrument report, at nominal time t, the level actually reached at $t+\delta$, so that $\eta_{\mathrm{obs}}(t) = \eta(t+\delta) \simeq \eta(t) + \delta\,\mathrm{d}\eta/\mathrm{d}t$. Because the rate of change of the total sea level is dominated by the tide, subtracting the correctly dated harmonic prediction η_{tide} leaves a surge of the form

$$s_{\mathrm{obs}}(t) \simeq s_{\mathrm{real}}(t) + \delta\,\frac{\mathrm{d}\eta_{\mathrm{tide}}}{\mathrm{d}t}, \tag{5}$$

that is, a timing error injects into the surge a scaled copy of the rate of change of the tide, a signal that is known exactly. We therefore estimate δ in each 30-day window (stepped by 15 days, and requiring at least 60 % of good observations) by least squares regression of the surge on $\mathrm{d}\eta_{\mathrm{tide}}/\mathrm{d}t$ without intercept, following the approach used to characterise tide gauge

https://doi.org/10.5194/egusphere-2026-5073
Preprint. Discussion started: 31 August 2026

timing errors (Agnew, 1986; Martín Míguez et al., 2008; Hudson et al., 2017). The real surge is meteorological and carries no preferred tidal phase, so its projection averages out over the window, whereas the leakage projects systematically and survives.

The estimated offset is then converted into the amplitude of the spurious surge it injects, $A = |\delta|\,\sigma(\mathrm{d}\eta_{\mathrm{tide}}/\mathrm{d}t)$, expressed in metres. This conversion is what makes the diagnostic portable: the same offset of a few minutes injects a 20 cm surge at a macrotidal station such as Calais, where $\sigma(\mathrm{d}\eta_{\mathrm{tide}}/\mathrm{d}t) = 0.93\,\mathrm{m\,h^{-1}}$, and essentially nothing where the tide is small, in which case δ is in addition estimated very poorly. Expressed in metres the leakage is directly comparable with the surge itself, whose standard deviation at Calais is 0.234 m.

The same quantity is computed for the MAROR reconstruction, which provides the reference level of the diagnostic. Being forcing-driven, MAROR cannot manufacture tidal leakage through a timing error, but it does inherit whatever small residual the harmonic prediction leaves in any realistic sea-level series, so its own amplitude A_{mod} measures the noise floor of the estimator at that station and period. Only leakage that the observation does not share with the reconstruction is attributable to the record. A window is therefore flagged when the excess leakage exceeds 0.05 m, $|\delta_{\mathrm{obs}} - \delta_{\mathrm{mod}}|\,\sigma(\mathrm{d}\eta_{\mathrm{tide}}/\mathrm{d}t) > 0.05$ m, and the observed leakage is at the same time at least 1.5 times the modelled one, the second condition discarding windows in which observation and reconstruction share the same residual. Every hour covered by a flagged window is discarded. Figure 6b shows that the detector is activated over the suspect period and stays silent elsewhere. Across the North Sea the incidence of flagged data falls monotonically with time, from $6.9\,\%$ of the hours in the 1970s to $0.3\,\%$ in the 2010s, consistent with the gradual replacement of mechanical instruments by digital ones.

The value of the diagnostic is that this defect was invisible to conventional quality control performed in this study. The Calais record has no impossible values, no datum jumps and no gaps; every value is individually plausible, and the fault lies only in the phase of the record. Removing the flagged stretches raises the test correlation at Calais from 0.62 over all test hours to 0.83 over the surviving hours (Fig. 6a). We further retrained the North Sea network after applying the detector, in one experiment only to Calais and in another to the eight tide gauges with more than $5\,\%$ of their hours flagged (Calais, Lauwersoog, Schiermonnikoog, Nieuwe Statenzijl, Oscarsborg, Nes, Immingham and Sheerness), to test whether removing the corrupted data during training improves the reconstruction. Evaluated over the same surviving hours, the retrained correlation at Calais moves only marginally, to 0.82 and 0.86 respectively, against the 0.83 already obtained by scoring the original reconstruction on those hours, and the reconstructions themselves shift by only a few centimetres. Retraining therefore brings no material improvement, and it is enough to detect and discard the flagged data.

Two caveats bound the scope of the detector. First, it is sensitive only to phase differences between the tide and the surge, which is the fingerprint of a phase error; an amplitude error, such as a mis-scaled tide gauge or a biased tidal prediction, is in phase with the tide and would not be seen, and the detector likewise does not react to datum jumps or slow drifts, which would require separate checks. Second, a genuine oscillation near the semidiurnal band, such as a tidally forced harbour seiche, could in principle be tide-coherent as well. What protects the interpretation here is that the detector relies on coherence with the phase of $\mathrm{d}\eta_{\mathrm{tide}}/\mathrm{d}t$ rather than on energy at a fixed frequency, that the signal appears and disappears in time rather than being stationary as a geometric resonance would be, and that the neighbouring tide gauges do not share it. The broader message is

https://doi.org/10.5194/egusphere-2026-5073
Preprint. Discussion started: 31 August 2026

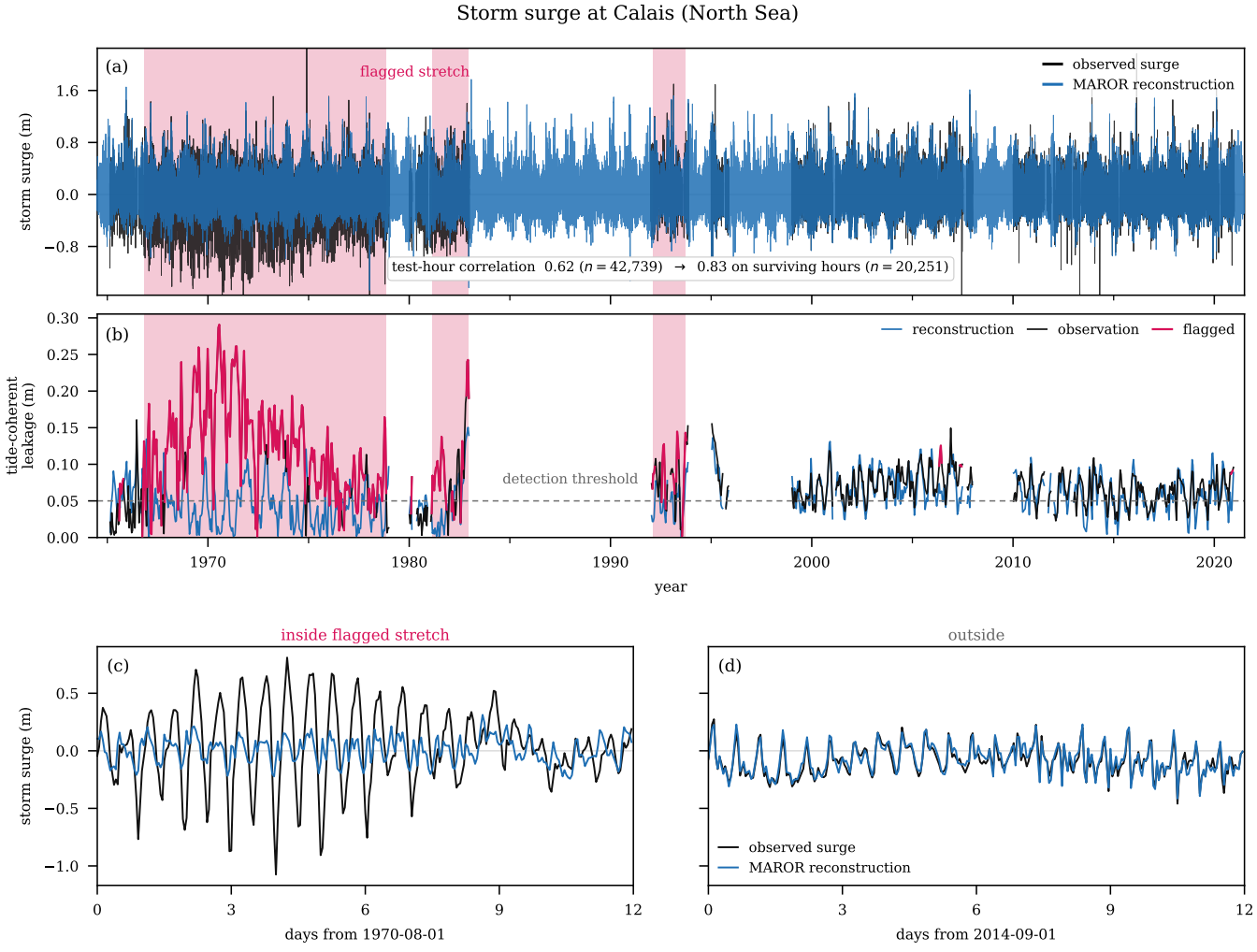

Figure 6. MAROR as a data-quality diagnostic, illustrated at Calais (North Sea). (a) De-tided storm surge over the full record: observation (black) and MAROR reconstruction (blue), with the periods most affected shaded as in (b). The observed surge carries an excess of variance between the late 1960s and the early 1980s that the reconstruction does not reproduce; removing the flagged hours raises the test-year correlation from 0.62 over all test hours to 0.83 over the surviving hours. (b) Detector output: the amplitude of the tide-coherent leakage (in metres) in the observation (black) and in the reconstruction (blue), with the observed leakage over the flagged 30-day windows drawn in red. A window is flagged, and its hours discarded, when the observed leakage exceeds the reconstructed one by more than the 0.05 m threshold (dashed) and is at the same time at least 1.5 times as large. The shading spans the periods most affected and is repeated in (a) as a visual guide; the hours actually discarded are those of the flagged windows, drawn in red. (c, d) Two 12-day windows of the de-tided surge, one inside a flagged stretch (c) and one outside it (d). Inside, the observation oscillates at tidal frequencies while the forcing-driven reconstruction stays flat, the signature of a tide gauge clock error; outside, the two agree.

https://doi.org/10.5194/egusphere-2026-5073
Preprint. Discussion started: 31 August 2026

methodological: MAROR can be used in two steps, first training the network to reconstruct the surge, and then using it as a reference to flag, and either discard or correct, the stretches of a record that are out of phase with the physics.

5 Discussion and Conclusions

MAROR reconstructs gap-free, hourly coastal sea level that is based on real tide gauge observations and driven by atmospheric forcing, SST, waves, and, where available, altimetric forcing. It extends the HIDRA3 architecture (Rus et al., 2025) in three ways that together turn a short-range forecast model into a multi-decadal sea level reconstruction tool: a long-range autoregressive rollout with an exposure-bias correction that removes the drift such rollouts would otherwise accumulate (Sect. 2.4); the use of satellite altimetry, so that the network can capture the slow ocean variability that atmospheric reanalysis alone misses; and a generative ensemble that delivers a per-tide-gauge, per-lead-time uncertainty band, calibrated by a light post-hoc inflation. Being at once based on observations, forcing-driven, and uncertainty-calibrated is what sets MAROR apart from purely hydrodynamic hindcasts, which are physically consistent but not tied to the local record, and from earlier data-driven methods, which do not provide a calibrated predictive distribution. Applied to the 373 tide gauges of the eleven study regions, MAROR reconstructs the hourly storm surge with a median correlation of 0.88 and a median skill score of 0.52 (Sect. 4.1), continuously from 1940 to the present.

Comparison with existing products. To place this performance in context, we compared MAROR against two reference products of different nature: a physically based hydrodynamic hindcast (the GTSM-ERA5 reanalysis; Muis et al. 2020) and a modern deep-learning model (Tiggeloven et al., 2021). Both comparisons are carried out on the de-tided storm surge. The two comparisons cover different subsets of the 373 tide gauges, for reasons intrinsic to each reference product. GTSM-ERA5 is a gridded hydrodynamic product, so the comparison is restricted to the 326 tide gauges that have at least one GTSM output point within 10 km with concurrent data; if more than one point is available, the modelled surge is averaged. The excluded gauges sit in estuaries or inner bays that the model grid does not resolve. The deep-learning model, in turn, is published only at its own set of stations, of which 171 could be matched to our tide gauges and had overlapping data. Against the hydrodynamic GTSM-ERA5, MAROR lowers the median surge RMSE from 0.10 to 0.05 m and raises the median correlation from 0.71 to 0.94, outperforming it at 326 of 326 tide gauges. Compared to the deep-learning model of Tiggeloven et al. (2021), MAROR is better at 151 of 171 tide gauges, with a median RMSE of 0.04 against 0.07 m and a median correlation of 0.96 against 0.78. Two caveats bound this last comparison. First, the published predictions of Tiggeloven et al. (2021) cover only their one-year held-out test period at each station, so the comparison is made over those hours and each model being scored against its own observations; second, where that year falls within the MAROR training set the comparison favours MAROR, and we therefore report it separately for the subset of 24 tide gauges whose window is also held out by MAROR, where MAROR remains better at 22 of 24, with a median RMSE of 0.08 against 0.16 m and a median correlation of 0.96 against 0.54. These stations fall in the MAROR test years, which are balanced by extremes across the record (Sect. 3.4) and therefore sample the most severe events; both models consequently show a larger RMSE here than over the full set, simply because the surges are larger, but MAROR's median correlation is essentially unchanged from the full set (0.96), whereas the deep-learning baseline degrades markedly on

https://doi.org/10.5194/egusphere-2026-5073
Preprint. Discussion started: 31 August 2026

these extreme-selected years (from 0.78 to 0.54). Overall, MAROR clearly outperforms the physical hindcast and is competitive with, and on average better than, the state-of-the-art deep-learning baseline, while adding a calibrated uncertainty.

Two distinct causes of low skill. One of the results is that a low reconstruction skill can have two very different origins, and that MAROR helps to differentiate them. The first is physical: where a large part of the surge variance is set by processes the atmospheric forcing does not resolve, such as the compact cores of tropical cyclones (Sect. 4.3) or the slow ocean dynamics of western boundary currents (Sect. 4.2), the reconstruction is limited by the forcing rather than by the network, and no amount of training can recover a signal that is not present in the inputs. The satellite-altimetry channel, included in MAROR, is introduced to supply the slow ocean variability, such as that of the western boundary currents, that the atmospheric forcing lacks, and it proves both effective and, in fact, necessary where that oceanic signal dominates. This is most clearly demonstrated in southern Japan, where adding the altimetry markedly improves the reconstruction of the Kuroshio-driven variability and is required to reach a better skill there (Sect. 4.2). The second is instrumental: where the forcing does resolve the surge and neighbouring tide gauges are reconstructed well, an anomalously low skill points instead to a defect in the observation. Because MAROR is trained jointly on many tide gauges and is physically plausible by construction, it provides an independent reference against which each record can be checked, turning the reconstruction into a data-quality diagnostic (Sect. 4.4) that flags, for instance, tide-gauge clock errors invisible to conventional quality control. This ability to separate a genuine physical limit from a data problem is, in our view, as valuable as the reconstruction itself.

Extending records for extreme-value analysis. Because MAROR produces a homogeneous, gap-free hourly surge time series over the full 1940-to-present period, it is directly useful for the statistics of extremes, where short records estimate long return periods poorly. Fitting the extreme-value distribution to the reconstructed annual maxima narrows the confidence bands and stabilises the long return periods relative to the few decades of observations (Sect. 4.2), and the reconstruction recovers historical events absent from the observed sample, such as the hurricane-like storm of February 1941 on the Atlantic coast of Iberian Peninsula. A reconstruction that follows the tail of the surge distribution, and not merely its bulk, is what makes this extension trustworthy for coastal flood-risk and long-term change assessments.

Toward the final product. The skill reported here is obtained under a 70/15/15 partition that withholds a substantial fraction of each record for validation and testing, so that performance can be measured on years the network never saw. For the operational release of the reconstruction the network can instead be trained on almost the entire record, holding out only the few years needed for early stopping the training process.

Limitations. Several limitations should be kept in mind. The reconstruction inherits the resolution limits of its forcing: the atmospheric reanalysis underrepresents the intensity of tropical cyclones, so the most compact and extreme surge peaks are systematically, if modestly, underestimated (Sect. 4.3), a shortcoming shared by hydrodynamic models driven by the same fields and only partly remedied by blending in parametric cyclone winds. The earliest decades rest on a sparser observational basis, since ERA5 is of lower quality before the satellite era (Sect. 3.2), and should be interpreted with correspondingly lower confidence. A separate network is trained for each region rather than a single global model, a deliberate choice that keeps each domain small enough for the shared geophysical encoder but that does not yet exploit information across regions. The network reconstructs sea level only at trained tide gauges and is therefore not spatially complete, unlike a hydrodynamic hindcast;

https://doi.org/10.5194/egusphere-2026-5073
Preprint. Discussion started: 31 August 2026

making its location-specific modules dependent on site descriptors would extend it to any coastal point. The altimetry, as ingested here, does not fully recover the oceanic low-frequency variance even where it helps most.

Code and data availability. An interactive companion to this paper is openly available at https://maror.angelamores.org, where the reconstructed storm surge can be inspected against the observations at each of the 373 tide gauges over the full 1940 to 2026 period, together with the corresponding skill metrics, Taylor diagrams and quantile–quantile plots. The model code is hosted at https://github.com/DrAmores/MAROR, currently under restricted access and to be made public upon publication.

Author contributions. A. Amores designed the MAROR neural network and the conceptual approach of the study, performed the computations and the analysis, and wrote the manuscript. M. Marcos contributed to the conceptual design of the study and to the writing. R. Alcaraz contributed to the design of the neural network. A. Martín contributed to the quality control of the tide gauge data. P. Thompson and T. Wahl contributed to the discussion of the results. All authors contributed to the writing of the final version of the manuscript.

Competing interests. The authors declare that they have no competing interests.

Acknowledgements. The authors gratefully acknowledge the supercomputing resources provided by the Balearic Islands Centre for Supercomputing and AI (BSAI). The MAROR networks were trained and run on the TalaIA supercomputer.

The authors used Claude (Anthropic) to assist with editing and formatting of the manuscript text and figure captions. Claude also assisted developing the MAROR's code.

Financial support. This research was conducted under the SeaLine project (reference PID2021-124085OB-I00) funded by MICIU/AEI /10.13039/501100011033 and by FEDER, UE.

https://doi.org/10.5194/egusphere-2026-5073
Preprint. Discussion started: 31 August 2026

References

Agnew, D. C.: Detailed analysis of tide gauge data: A case history, Marine Geodesy, 10, 231–255, https://doi.org/10.1080/01490418609388024, 1986.

Agulles, M., Marcos, M., Amores, Á., and Toomey, T.: Storm surge modelling along European coastlines: The effect of the spatio-temporal resolution of the atmospheric forcing, Ocean Modelling, 192, 102 432, https://doi.org/10.1016/j.ocemod.2024.102432, 2024.

Bengio, S., Vinyals, O., Jaitly, N., and Shazeer, N.: Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks, in: Advances in Neural Information Processing Systems, edited by Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., vol. 28, Curran Associates, Inc., https://proceedings.neurips.cc/paper_files/paper/2015/file/e995f98d56967d946471af29d7bf99f1-Paper.pdf, 2015.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q.: Accurate medium-range global weather forecasting with 3D neural networks, Nature, 619, 533–538, https://doi.org/10.1038/s41586-023-06185-3, 2023.

Camps-Valls, G., Fernández-Torres, M.-Á., Cohrs, K., Höhl, A., Castelletti, A., et al.: Artificial intelligence for modeling and understanding extreme weather and climate events, Nature Communications, 16, 1919, https://doi.org/10.1038/s41467-025-56573-8, 2025.

Cid, A., Camus, P., Castanedo, S., Méndez, F. J., and Medina, R.: Global reconstructed daily surge levels from the 20th Century Reanalysis (1871–2010), Global and Planetary Change, 148, 9–21, https://doi.org/10.1016/j.gloplacha.2016.11.006, 2017.

Codiga, D. L.: Unified Tidal Analysis and Prediction Using the UTide Matlab Functions, Tech. Rep. 2011-01, Graduate School of Oceanography, University of Rhode Island, Narragansett, RI, 2011.

Dullaart, J. C. M., Muis, S., Bloemendaal, N., and Aerts, J. C. J. H.: Advancing global storm surge modelling using the new ERA5 climate reanalysis, Climate Dynamics, 54, 1007–1021, https://doi.org/10.1007/s00382-019-05044-0, 2020.

E.U. Copernicus Marine Service Information (CMEMS): Global Ocean Gridded L4 Sea Surface Heights and Derived Variables Reprocessed (SSALTO/DUACS Delayed-Time, DT2024 baseline, 1/8°), Marine Data Store (MDS), https://doi.org/10.48670/moi-00148, product SEALEVEL_GLO_PHY_L4_MY_008_047, DT2024 0.125 degree daily dataset, version vNov2024; accessed 19 June 2025, 2024.

Ferro, C. A. T.: Fair scores for ensemble forecasts, Quarterly Journal of the Royal Meteorological Society, 140, 1917–1923, https://doi.org/10.1002/qj.2270, 2014.

Gneiting, T. and Raftery, A. E.: Strictly proper scoring rules, prediction, and estimation, Journal of the American Statistical Association, 102, 359–378, https://doi.org/10.1198/016214506000001437, 2007.

Haigh, I. D., Marcos, M., Talke, S. A., Woodworth, P. L., Hunter, J. R., Hague, B. S., Arns, A., Bradshaw, E., and Thompson, P.: GESLA Version 3: A major update to the global higher-frequency sea-level dataset, Geoscience Data Journal, 10, 293–314, https://doi.org/10.1002/gdj3.174, 2023.

Ham, Y.-G., Kim, J.-H., and Luo, J.-J.: Deep learning for multi-year ENSO forecasts, Nature, 573, 568–572, https://doi.org/10.1038/s41586-019-1559-7, 2019.

He, K., Zhang, X., Ren, S., and Sun, J.: Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, https://doi.org/10.1109/CVPR.2016.90, 2016.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Vil-

https://doi.org/10.5194/egusphere-2026-5073
Preprint. Discussion started: 31 August 2026

laume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Hudson, A. S., Talke, S. A., Branch, R., Chickadel, C., Farquharson, G., and Jessup, A.: Remote measurements of tides and river slope using an airborne lidar instrument, Journal of Atmospheric and Oceanic Technology, 34, 897–904, https://doi.org/10.1175/JTECH-D-16-0197.1, 2017.

Killick, R., Fearnhead, P., and Eckley, I. A.: Optimal detection of changepoints with a linear computational cost, Journal of the American Statistical Association, 107, 1590–1598, https://doi.org/10.1080/01621459.2012.737745, 2012.

Klambauer, G., Unterthiner, T., Mayr, A., and Hochreiter, S.: Self-Normalizing Neural Networks, in: Advances in Neural Information Processing Systems, edited by Guyon, I., von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., vol. 30, pp. 971–980, Curran Associates, Inc., https://proceedings.neurips.cc/paper_files/paper/2017/hash/5d44ee6f2c3f71b73125876103c8f6c4-Abstract.html, 2017.

Lam, R., Sanchez-Gonzalez, A., Willson, M., et al.: Learning skillful medium-range global weather forecasting, Science, 382, 1416–1421, https://doi.org/10.1126/science.adi2336, 2023.

Lang, S., Alexe, M., Clare, M. C. A., Roberts, C., Adewoyin, R., Ben Bouallègue, Z., Chantry, M., Dramsch, J., Dueben, P. D., Hahner, S., Maciel, P., Prieto-Nemesio, A., O'Brien, C., Pinault, F., Polster, J., Raoult, B., Tietsche, S., and Leutbecher, M.: AIFS-CRPS: ensemble forecasting using a model trained with a loss function based on the continuous ranked probability score, npj Artificial Intelligence, 2, 18, https://doi.org/10.1038/s44387-026-00073-7, 2026.

Martín Míguez, B., Testut, L., and Wöppelmann, G.: The Van de Casteele test revisited: An efficient approach to tide gauge error characterization, Journal of Atmospheric and Oceanic Technology, 25, 1238–1244, https://doi.org/10.1175/2007JTECHO554.1, 2008.

Materia, S. et al.: Artificial intelligence for climate prediction of extremes: State of the art, challenges, and future perspectives, WIREs Climate Change, 15, e914, https://doi.org/10.1002/wcc.914, 2024.

Muis, S., Verlaan, M., Winsemius, H. C., Aerts, J. C. J. H., and Ward, P. J.: A global reanalysis of storm surges and extreme sea levels, Nature Communications, 7, 11 969, https://doi.org/10.1038/ncomms11969, 2016.

Muis, S., Apecechea, M. I., Dullaart, J., de Lima Rego, J., Madsen, K. S., Su, J., Yan, K., and Verlaan, M.: A high-resolution global dataset of extreme sea levels, tides, and storm surges, including future projections, Frontiers in Marine Science, 7, 263, https://doi.org/10.3389/fmars.2020.00263, 2020.

Nagaraj, M., Rodriguez, A., and Wahl, T.: Regional Modeling of Storm Surges Using Localized Features and Transfer Learning, Journal of Geophysical Research: Machine Learning and Computation, 2, e2025JH000 650, https://doi.org/10.1029/2025JH000650, 2025.

Rus, M., Fettich, A., Kristan, M., and Ličer, M.: HIDRA2: deep-learning ensemble sea level and storm tide forecasting in the presence of seiches – the case of the northern Adriatic, Geoscientific Model Development, 16, 271–288, https://doi.org/10.5194/gmd-16-271-2023, 2023.

Rus, M., Mihanović, H., Ličer, M., and Kristan, M.: HIDRA3: a deep-learning model for multipoint ensemble sea level forecasting in the presence of tide gauge sensor failures, Geoscientific Model Development, 18, 605–620, 2025.

Taburet, G., Sanchez-Roman, A., Ballarotta, M., Pujol, M.-I., Legeais, J.-F., Fournier, F., Faugere, Y., and Dibarboure, G.: DUACS DT2018: 25 years of reprocessed sea level altimetry products, Ocean Science, 15, 1207–1224, https://doi.org/10.5194/os-15-1207-2019, 2019.

Tadesse, M., Wahl, T., and Cid, A.: Data-driven modeling of global storm surges, Frontiers in Marine Science, 7, 260, https://doi.org/10.3389/fmars.2020.00260, 2020.

https://doi.org/10.5194/egusphere-2026-5073
Preprint. Discussion started: 31 August 2026

Tiggeloven, T., Couasnon, A., van Straaten, C., Muis, S., and Ward, P. J.: Exploring deep learning capabilities for surge predictions in coastal areas, Scientific Reports, 11, 17 224, https://doi.org/10.1038/s41598-021-96674-0, 2021.

Žust, L., Fettich, A., Kristan, M., and Ličer, M.: HIDRA 1.0: deep-learning-based ensemble sea level forecasting in the northern Adriatic, Geoscientific Model Development, 14, 2057–2074, https://doi.org/10.5194/gmd-14-2057-2021, 2021.