Implicit Assimilation of Sparse In Situ Data
for Dense & Global Storm Surge Forecasting

Patrick Ebel[*]

patrick.ebel@esa.int

Brandon Victor[†]

b.victor@latrobe.edu.au

Peter Naylor[*]

peter.naylor@esa.int

Gabriele Meoni[*,°]

gabriele.meoni@esa.int

Federico Serva[‡]

federico.serva@terrarum.eu

Rochelle Schneider[*]

rochelle.schneider@esa.int

[*]European Space Agency, Φ-lab   [†] La Trobe University   [°]European Space Agency, ACT   [‡] Consiglio Nazionale delle Ricerche

Abstract

Hurricanes and coastal floods are among the most disastrous natural hazards. Both are intimately related to storm surges, as their causes and effects, respectively. However, the short-term forecasting of storm surges has proven challenging, especially when targeting previously unseen locations or sites without tidal gauges. Furthermore, recent work improved short and medium-term weather forecasting but the handling of raw unassimilated data remains nontrivial. In this paper, we tackle both challenges and demonstrate that neural networks can implicitly assimilate sparse in situ tide gauge data with coarse ocean state reanalysis in order to forecast storm surges. We curate a global dataset to learn and validate the dense prediction of storm surges, building on preceding efforts. Other than prior work limited to known gauges, our approach extends to ungauged sites, paving the way for global storm surge forecasting.

1. Introduction

Space-borne Earth observation allows for large-scale monitoring of our planet, its atmosphere and events such as natural hazards that may pose significant threat to human life. While the strengths of satellite imagery are its broad spatial extent, its spatio-temporal resolution is inferior compared to on-site measurements. In contrast, in situ sensors may provide (sub-)hourly recordings at highest accuracy, yet they are sparsely deployed and thus lack spatial coverage. Fusing both kinds of data at a global scale holds promises, but harmonizing the sensors in a manner suitable for neural networks to process is an open research direction. A well-established paradigm tackling this issue in the context of weather analysis is that of data assimilation [10, 13]. However, it is computationally costly and not easily approachable. In this work, we address the challenge
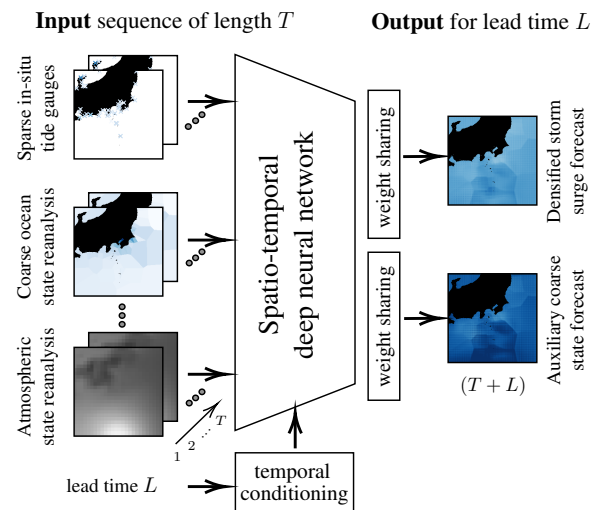

Figure 1. **Overview:** Our approach provides densified high-resolution storm surge forecasts (top) by implicitly assimilating inputs of sparse in situ tide gauge time series (top) with paired sequences of ocean (center) and weather state (bottom) re-analysis products. For additional supervision, coarse ocean state reanalysis maps (bottom) at coinciding lead time are also predicted.

of global storm surge forecasting by implicitly assimilating sparse and raw tide gauge data with coarse weather and ocean state reanalysis products. Storm surges are extreme weather-driven ocean dynamics superimposed on the mean sea level and tidal rhythms, which can cause coastal floods. Scientific consensus is that the coming decades will bring a sharp increase of coastal hazards due to climate change-caused rise in mean sea levels [17, 22, 36], aggravated by land subsidence [39] and more intense extreme storm events [2, 8]. Our work on storm surge forecasting is motivated to address such hazards, aligning with the *United Nation Sustainable Development Goals* 11.5 & 13 [28, 31]. Particularly, our approach is inspired by recent

advances in weather forecasting that allow generalization to previously unencountered or ungauged sites, which may especially benefit under-served communities with less access to well-maintained in situ measurement infrastructure. To empower worldwide AI-driven surge forecasting, we curate a novel global and multiple decades spanning dataset of in situ tidal gauge records, paired with weather and ocean state reanalysis, all preprocessed according to the best domain-specific practices. We highlight the dataset's worth by benchmarking a diverse landscape of approaches including conventional forecasting techniques, an operational numerical model, state-of-the-art deep neural networks, a recent vision transformer for weather forecasting and our enhanced adaptation of a popular lightweight temporal attention network. While we evince the competitiveness of the latter model, the main objective of our experiments is to demonstrate the predictability of storm surges at previously unencountered or altogether ungauged locations, an aim whose feasibility has been questioned in prior work [4, 37].

In sum, our main contributions are two-fold:

- We introduce a novel, global and multi-decadal dataset of in situ ocean surge time series, paired with atmospheric and ocean state reanalysis products, spuring and facilitating further research on this critical matter.
- We demonstrate that precise and dense storm surge forecasts can be obtained by fusing sparse in situ data of coastal tide gauges with coarse atmosphere and ocean reanalysis. Critically, our forecasts extend to previously unseen gauges and entirely ungauged locations, which may benefit under-served communities.

2. Related Work

2.1. Short-to-medium Range Weather Forecasting

Recent work marked notable progress on numerical weather prediction [3, 20, 26, 34]. Such models typically rely on atmospheric initial values from a reanalysis product such as ERA5 [10], i.e. best estimates derived by updating prior knowledge with multi-source weather observations. Contrarily, the contribution of [1] proposes an implicit assimilation approach to fuse in situ weather radar station data with coarse resolution reanalysis products, yielding dense and skillful rainfall forecasts over the United States. While we draw inspiration from this approach, our focus is on the global coastlines to model marine dynamics. Technically, our approach deviates from the aforementioned ones by processing local patches of data instead of a coarsely resolved global context in a single forward pass of the network, and is thus significantly more lightweight.

2.2. Storm Surge Forecasting

Operational storm surge forecasting pre-dates deep learning, with early techniques explicitly modeling the physics of maritime dynamics with a focus on particular ocean basins [21, 35]. Of particular interest is the Global Tide and Surge Model (GTSM), a hydrodynamic model forced with ERA5 to globally predict surge on an irregular grid. In this study, we built upon coarsely resolved GTSM ocean state analysis [23] to drive the assimilation of raw in situ data. That is, we fuse the reanalysis product with accurate but sparse in situ records for improved and densified surge predictions.

Initial efforts for global storm surge modeling via deep learning are given by [4, 37]. Both studies are limited to temporal generalization, i.e. they evaluate on gauges trained upon and solely generalize to future time points of these gauges. In contrast, our data and approach enable to generalize in both time and space: Relatedly, recent work [25] proved the feasibility of river streamflow predictions at ungauged basins, in the spirit of which we generalize coastal storm surge forecasting to unseen shores. This defeats the prevailing wisdom that ocean modeling necessitates at least 6-7 prior years of training data at any site of interest [4, 37].

3. Data

We collect a new global multi-decadal dataset combining co-registered atmosphere reanalysis, ocean state reanalysis and pre-processed in situ tide gauge measurements. All data is sampled to an hourly frequency with dates ranging from the beginning of 1979 to the start of 2019, and gridded at $0.025°$ spatial resolution. The atmosphere reanalysis denotes best estimates of historical mean sea level pressure as well as 10 metre U and V wind components ('msl', 'u10' and 'v10', respectively) at about 30 km resolution, as provided by the ERA5 catalogue [10]. The ocean-state model forced by ERA5 meteorology provides the storm surge residual based upon the irregularly gridded Deltares Global Tide and Surge Model (GTSM) forced via the aforementioned ERA5 inputs, as given by the Copernicus Climate Change Service (C3S) Climate Data Store (CDS) [23]. Furthermore, a global land-sea mask [15, 40] resampled to circa 3 km resolution is provided. Finally, precise storm surge measurements are derived from in situ tide gauge records collected in GESLA-3 [9] and spatially distributed as shown in Fig. 2. Principally, our pre-processing pipeline follows the established workflow of [37]; featuring mean sea level de-trending, harmonic decompositioning [5] and de-noising steps. Key differences to the prior work are their filtering of any in situ sites with records shorter than seven years, whereas we take inspiration from recent progress on (un)gauged river streamflow forecasting [25] and keep such data. While shorter durations pose a greater challenge to learn site-wise dynamics, this drastically increases the overall amount of valuable in situ data from a total of 736 tidal gauges in [37] to 3553 locations in our work — yielding an almost five-fold increase of valuable in situ data compared to preceding efforts.
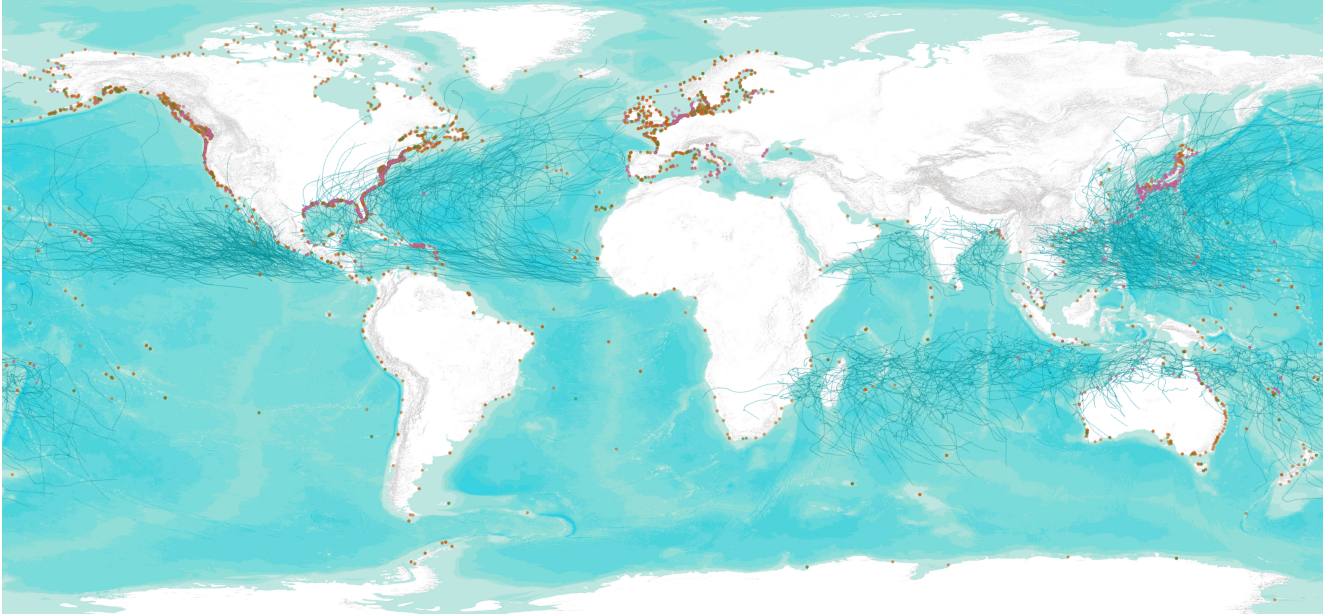
Figure 2. **Data.** Green & orange dots denote storm surge time series locations with records in 1979-2019, as pre-processed from the GESLA-3 collection of tide gauges [9]. Dark lines indicate hurricane tracks in 2014-2019 as indexed by IBTRaCS [18]. Pink markers highlight test split gauges, biased to points of landfall. Visualizations of the ERA5 grid and the irregular GTSM grid are omitted for brevity.

4. Methods

The problem tackled herein is that of forecasting the highly non-linear dynamics of storm surges on a short lead time: Every sample $i = 0, 1, ..., |\mathcal{D}|$ of the dataset $\mathcal{D}$ denotes a pair $(\boldsymbol{X^i}, Y^i)$, with $\boldsymbol{X}^i = [X_1^i, \cdots, X_T^i]$ being the input time series of size $[T \times C_{in} \times H \times W]$ featuring in situ plus atmospheric reanalysis and model-based surge data, and Y^i is the target image of shape $[C_{out} \times H \times W]$ at a lead time of L hours. T is the temporal length of the input series, C_{in} and C_{out} denote the number of input and output channels, and $H \times W$ the images' two spatial dimensions. For convenience, the i superscript is omitted in the remainder of the paper. Unless stated otherwise, we set $T = 12$, $C_{in} = 5$, $C_{out} = 2$, $H = W = 256$ and $L = 8$ hours, which has been a common choice in prior short-term forecasting works [34]. Example data are illustrated in Fig. 3, showcasing the time series dynamics, the diversity in context and the presence of extreme weather events in the data.

4.1. Models

We demonstrate the feasibility of our implicit assimilation and densification approach by adapting a representative variety of models. Besides highlighting our paradigm's effectiveness for sparse coastal observations, these evaluations may serve as a benchmark for future research. For each considered network, we follow the respective architecture's best practice in terms of hyperparameters given in the referenced literature, unless otherwise specified. Models are:

Conventional baselines As a simple baseline, we consider the seasonal average surge based on historic values at the gauge of interest and the given target time. This necessitates historical data at the target gauge, but is expected to provide a solid baseline for sites experiencing seasonal cyclone activity [32]. Second, we consider the mean surge at the gauge of interest over the input time period. While again requiring access to the gauge's records, this would prove beneficial whenever the surge at target time doesn't deviate too much from the input period's. Third, we consider the linear extrapolation of surge time series inputs to the target time. Finally, the global physical storm surge predictions of GTSM forced with ERA5 data are reported and compared against [23]. As GTSM is numerically simulated on an irregular grid [6, 14] we perform nearest neighbor extrapolation to get coarse forecasts at any site of interest, specifically for getting predictions at unknown test split gauges, and then extrapolate these to the target time.

LSTM-based Long short-term memory (LSTM) networks [11] are the state-of-the-art architectures for temporal modeling of fluvial [19, 25] and coastal dynamics [37]. Accordingly, we consider classical as well as convolutional LSTM (ConvLSTM) [33] for storm surge forecasting. We use the architectures from [37] and minimally adapt them for the sake of comparability to accept our more comprehensive data, including time series of preceding gauge values. As such, both models receive a 5x5 context window around the gauge, and predict a single point in the future.

	Gauges	ERA5 msl	ERA5 u10	ERA5 v10	GTSM	Gauges	GTSM

(a) In situ time series.　　　　　　　　　　　　　　(b) Input　　　　　　　　　　　　　　(c) Target
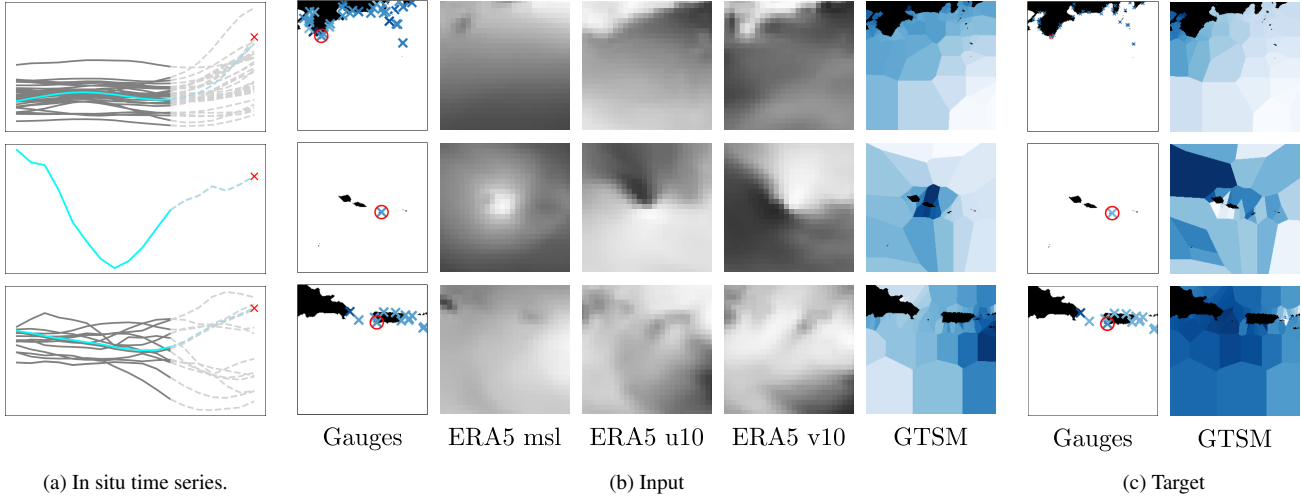
Figure 3. **Example data**, one local sample per row. Input grouped at the left, targets at the right. Inputs: Time series of target (blue) and context gauges (grey), with target history only given in the hyperlocal setting. gauge locations and surge values. ERA5 pressure at mean sea level, wind at 10 m in u and v directions. Coarse GTSM input. Targets: Surge at target time, GTSM at target time.

Attention-based We consider spatio-temporal transformer models, which both principally share a common structure of inputs and outputs as depicted in Fig. 1. First, we evaluate a MaxVIT U-Net backbone [38] as recently proposed for weather forecasting in [1], adapted to our problem statement. The network collates temporal information into the channel dimension, but is conditioned on the lead time of the target via Feature-wise Linear Modulation (FiLM) [27]. Finally, we consider the U-TAE of [7], originally proposed for panoptic segmentation. We adjust U-TAE by introducing FiLM at each of its convolution blocks—such that additionally to temporal embeddings at the input time points, our adaptation of the model is conditioned to forecast at a variable target time. Notably, the key difference between the last two models is that [38] resolves the input time series into the channel dimension and doesn't model time dynamics explicitly, whereas [7] processes time series explicitly and applies lightweight temporal attention but doesn't explicitly model global spatial interactions.

4.2. Densification

Central to our approach generalizing storm surge forecasting to previously unseen or ungauged sites is the concept of densification. For any model that outputs a two-dimensional storm surge forecast map $\hat{\boldsymbol{y}}_s$, we implement densification via the built-in spatial parameter sharing of the convolution operator. Specifically, we utilize 1×1 convolution kernels with a stride of 1 at the final network layer to broadcast from sparsely populated to non-observed pixel coordinates in the spatial dimensions. Auxiliary supervision and input data dropout are also used to further encourage the networks to learn densification, as proposed by Andrychowicz et al [1].

Auxiliary supervision Complementarily to learning a densified forecast of the sparse in situ time series, the networks additionally predict a forecast $\hat{\boldsymbol{y}}_c$ of the coarse GTSM ocean state at the same lead time L, as depicted in Fig. 1. This way, the models receive additional feedback at pixels which would otherwise not be populated and the preceding shared layers internalize to implicitly assimilate the sparse observations with the coarse reanalysis. To only evaluate the coarse ocean state predictions over valid locations we mask the loss computation with a land sea mask $\mathbb{1}_{lsm}$.

In situ dropout To furthermore encourage the densifying networks to predict non-trivial outputs at unpopulated pixels within the sparse input time series we perform data dropout. Specifically, we randomly remove in situ tide gauges from the input with a probability p but keep all sites within the target patch, such that the network is forced to learn extrapolating to the dropped sites. We set $p = 0.25$ and include a binary validity mask $\mathbb{1}_{val}$ in the network inputs as proposed in a weather prediction context by [1].

In sum, the densifying network architectures output two maps $\hat{\boldsymbol{y}} = [\hat{\boldsymbol{y}}_s, \hat{\boldsymbol{y}}_c]$. Map $\hat{\boldsymbol{y}}_s$ densely predicts the sparse GESLA gauges, and $\hat{\boldsymbol{y}}_c$ predicts the spatially interpolated future coarse GTSM values. Thus, they are trained via a weighted combination of two masked L1 cost functions

$$\mathcal{L}_s(\hat{\boldsymbol{y}}_s, \boldsymbol{y}_{GESLA}) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{val}(j) \cdot \|\hat{\boldsymbol{y}}_j - \boldsymbol{y}_j\|_1 , \quad (1)$$

$$\mathcal{L}_c(\hat{\boldsymbol{y}}_c, \boldsymbol{y}_{GTSM}) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{lsm}(j) \cdot \|\hat{\boldsymbol{y}}_j - \boldsymbol{y}_j\|_1 , \quad (2)$$

masked via $\mathbb{1}_{val}$ and $\mathbb{1}_{lsm}$, resulting in the combined loss

$$\mathcal{L}(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \mathcal{L}_s(\hat{\boldsymbol{y}}_s, \boldsymbol{y}_{GESLA}) + \lambda \mathcal{L}_c(\hat{\boldsymbol{y}}_c, \boldsymbol{y}_{GTSM}) \; . \quad (3)$$

We set the hyperparameter $\lambda = \frac{1}{100}$, to account for the sparseness of in situ data in comparison to the pixel-wise coarse evaluations and to compensate for the resulting magnitudes of differences in supervision frequency across both domains as well as their respective loss terms.

4.3. Lead time conditioning

To enable a flexible forecasting, accommodating for varying hours of look-ahead predictions at inference time and via a single forward-pass, we implement lead time conditioning via Feature-wise Linear Modulation (FiLM) [27]. Specifically, a shallow encoder projects queried lead times L into a low-dimensional feature space and linearly modulates convolutional feature maps via a learned scale and bias offset. We utilize lead time conditioning for the two considered densifying networks, and have one shallow encoder per each of their U-Net backbone's convolutional blocks.

5. Experiments

Splits   For our experiments, we set up splits by defining holdout data in terms of both the spatial and temporal dimensions: A globally distributed 20 % of coastal gauges across all ocean basins are reserved for the test split, whose temporal extent starts from April 2014 and is thus well observed via satellites, in the interest of follow-up works. The remaining 70 and 10 % of locations are utilized as training and validation splits, with records ranging from 1979 to 2014. This amounts to a total of 2561, 284 and 708 gauges for our train, validation and test split, respectively. The resulting map of in situ recordings (color-coded according to their splits) and hurricane trajectories is depicted in Fig. 2. The distribution of accessible and gauged sites is biased towards developed countries, underlining the need for machine learning solutions to serve under-represented regions.

Sampling   Having assigned a subset of gauges to the holdout splits, we determine the test dates by identifying coincidences of in situ records with storm tracks as given by IBTrACS [18]. When a storm passes within 100 km of a holdout tide gauge then we set such a date as the sample's target time. If no storm tracks pass by, then we resort to sampling target times at outlier surge values deviating more than 2 train split standard deviations from the train split's mean surge. At train time, we likewise perform outlier sampling with a probability of 0.5, roughly mirroring the distribution of (non) storm events at holdout gauges in the validation and test splits. If no such sample exists for the current gauge, then a random target date is drawn instead.
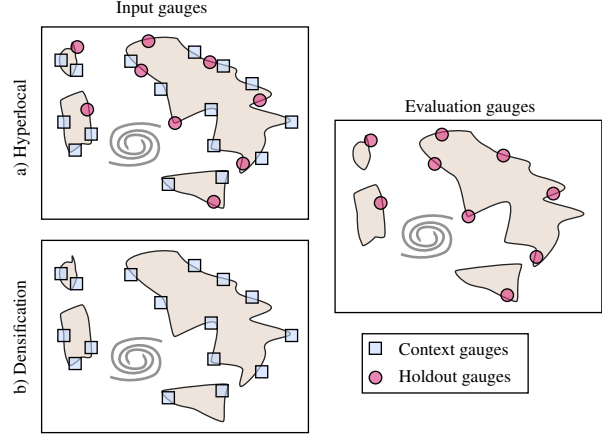

Figure 4. **Experimental setup.** Design of the densification and hyperlocal evaluation schemes, conceptualizing their respective inputs and outputs. The hyperlocal protocol focuses on forecasting of novel dynamics encountered at inference time, predicting surge at holdout target (green) and context gauges (blue) L hours ahead. The generalization setup quantifies the goodness of models to broadcast predictions to ungauged locations, i.e. unknown gauges not contained in the input and solely used for evaluation.

5.1. Implementation details

To enable fast online sampling of data and efficient training of deep networks, we represent all spatio-temporal data as netCDF files [29] via xarray [12], either loaded directly into memory or read in parallel via dask [30]. This is especially critical for the in situ GESLA-3 records, which are re-processed into a compact format as part of the preprocessing pipeline and get released with this publication.

In contrast to weather forecasting models [3, 20, 26] that process a global context window at proximately 30 km resolution, the networks we consider are significantly less resource-demanding and digest local patches of $(256 \, \text{px})^2$ at a finer pixel-resolution of circa 3 km to capture local variations of surge. All data features are z-standardized via their sufficient statistics calculated on the training split.

Training   For training, 1 epoch is defined by iterating over all train split gauges in a random order. At each gauge, a Gaussian is drawn around it's location to randomly sample what is treated as the local patches centroid c. The target date & time t are drawn randomly from the current gauge's records. Next, a lead time L is drawn randomly from $\{0, 1, 2, ..., 12\}$, as in [1]. The hourly input time sequence of length T is then given by the time interval of $[t - L - T, t - L]$. Note that the random sampling of c, t and L effectively acts as data augmentation. For autoregressive methods we evaluate the prediction at time t, whereas single-forward-pass approaches based on FiLM are conditioned on L to directly generate a forecast at time t.

Table 1. **Experimental evaluation.** We benchmark forecasting skills for $T = 12$ and $L = 8$ on a global holdout set of previously unseen gauges. Point-wise storm surge predictions at known gauges are evaluated in the hyperlocal setting (left), whereas the densification protocol (right) tests predictions at unknown or ungauged sites. FiLM U-TAE is the most competitive approach, followed by MaxVIT U-Net.

Model	Hyperlocal			Densification		
	↓ MAE (std)	↓ MSE (std)	↑ NNSE	↓ MAE (std)	↓ MSE (std)	↑ NNSE
seasonal average	0.281 (0.313)	0.177 (0.539)	0.424	—	—	—
input average	0.267 (0.295)	0.158 (0.452)	0.452	—	—	—
input extrapolation	0.182 (0.239)	0.090 (0.342)	0.593	—	—	—
GTSM extrapolation [23]	—	—	—	0.351 (0.643)	0.536 (4.744)	0.195
LSTM [11, 37]	0.166 (0.282)	0.107 (0.759)	0.595	—	—	—
ConvLSTM [33, 37]	0.162 (0.267)	0.098 (0.691)	0.618	—	—	—
FiLM U-TAE [7, 27]	**0.158 (0.209)**	**0.069 (0.248)**	**0.655**	0.190 (0.260)	**0.104 (0.535)**	**0.556**
MaxVIT U-Net [1, 38]	0.160 (0.212)	0.070 (0.263)	0.649	**0.178 (0.273)**	0.106 (0.587)	0.552

We use the ADAM optimizer [16] at a batch size of 16, with initial learning rates tuned over magnitudes 10^{-1} to 10^{-4} for each model individually. All networks train for 50 epochs with an exponential learning rate decay of 0.9. Models are evaluated on the validation split each epoch and the checkpoint with best validation loss is used for testing.

5.2. Evaluation

All network predictions at target time are compared against their respective test split in situ tide gauge values. Prediction goodness is evaluated in terms of Mean Absolute Error (MAE) as well as Mean Squared Error (MSE), reported in units of meters and with error-wise standard deviations (std) across the set of test split gauges denoted in brackets. Finally, we report each method's Normalized Nash-Sutcliffe Efficiency (NNSE) [24], which conceptually relates to the coefficient of determination (R^2) and takes values within 0 and 1. Similar to prior weather forecasting work [1] we evaluate in two experimental setups. The concepts of both setups are depicted in Fig. 4 and given as follows:

i. Hyperlocal evaluation In this experimental paradigm, the time series of holdout gauges not trained upon are included into the model's inputs at test time. Therefore, the challenge becomes to integrate previously unseen dynamics at novel locations and to assimilate newly encountered tidal gauges at inference time.

ii. Densification evaluation Predictions at previously unseen test split gauges are obtained via densification, i.e. a model's ability to predict at unknown locations is quantified here. Importantly, test split gauges are not part of the input time series and only used as targets. Note that this setup can't be accomplished via previous established approaches for storm surge forecasting not implementing densification, e.g. the conventional baselines and LSTM-based models.

6. Results

6.1. Main experiments

To demonstrate the feasibility of our problem statement and the benefits of our curated dataset, we evaluate all considered approaches according to their applicability in the hyperlocal and densification experimental schemes. Outcomes are reported in Table 1. The results show that FiLM U-TAE performs best in the hyperlocal setting, forecasting surge at newly encountered gauges with a mean absolute accuracy of circa 16 cm. The seasonal average and input average predictions are more erroneous, particularly in terms of squared error—validating the presence of storm-driven extremes in the test data. Overall, all neural networks outperform competing approaches in the hyperlocal setting.

In the densification experiment, FiLM U-TAE performs best, closely followed by MaxVIT U-Net. Notably, both densifying models denote a substantial improvement over the GTSM baseline, whose prediction goodness exhibits elevated standard deviations across gauges.

Altogether, FiLM U-TAE tends to outperform MaxVIT U-Net, implying that temporal self-attention is of greater benefit than visual attention for our spatio-temporal forecasting task. To convey a better understanding of the spatial dependency of errors, Fig. 5 analyzes our best model's performance across the globe. The analysis shows that forecasting in the tropics is particularly hard, confirming that the curated benchmark provides a challenging problem.

6.2. Ablation experiments

In order to further investigate the challenges of short-term storm surge forecasting and explore which design choices determine the quality of predictions, we conduct a series of ablation studies following the densification experimental protocol. All ablations are run with the FiLM U-TAE network, which the preceding main experiments identified as the best performing backbone for our approach.
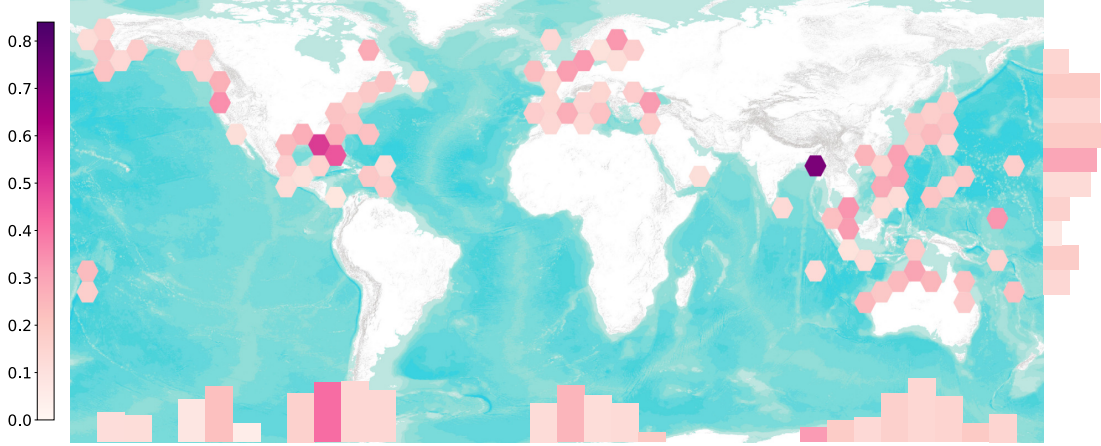
Figure 5. **Location** of holdout gauges impacts prediction performance. Average absolute errors in meters are color-coded and binned according to each sites' longitude and latitude coordinates, with gauge counts overlayed for each spatial dimension. Particularly challenging regions are the Gulf of Mexico, the Caribbean Sea and the Indian Ocean due to their extreme climate and resulting outsized surge dynamics.

Table 2. **Repeated Measures.** Evaluation of FiLM U-TAE with varying numbers of input time points T, flexibly accommodated for via temporal self-attention. Longer inputs tend to be beneficial.

input length T	↓ MAE (std)	↓ MSE (std)	↑ NNSE
6	0.194 (0.282)	0.115 (0.587)	0.551
12	0.190 (0.260)	0.104 (0.535)	0.556
18	**0.180 (0.230)**	**0.085 (0.510)**	**0.573**
24	**0.180 (0.230)**	**0.085 (0.510)**	0.571

Table 3. **Lead Time.** Evaluation of FiLM U-TAE with varying lead time offset L, modifiable thanks to lead time conditioning. Storm surge forecasts become more challenging the larger L gets.

lead time t	↓ MAE (std)	↓ MSE (std)	↑ NNSE
4	**0.169 (0.254)**	**0.093 (0.543)**	**0.583**
6	0.182 (0.269)	0.106 (0.551)	0.552
8	0.190 (0.260)	0.104 (0.535)	0.556
10	0.191 (0.273)	0.111 (0.553)	0.540
12	0.196 (0.273)	0.113 (0.539)	0.536

Table 4. **Input ablations.** Evaluation of our models with varying inputs. The outcomes underline the relevance of each modality.

input ablation	↓ MAE (std)	↓ MSE (std)	↑ NNSE
full model	0.190 (0.260)	**0.104 (0.535)**	**0.556**
no GTSM input	0.207 (0.284)	0.124 (0.543)	0.513
no ERA5 input	0.189 (0.273)	0.110 (0.545)	0.542
no data dropout	0.217 (0.289)	0.130 (0.539)	0.500
no FiLM, $L = 8$ fixed	**0.183 (0.273)**	0.108 (0.567)	0.547

Table 5. **Output ablations.** Evaluation of FiLM U-TAE with varying output channels. Ablations show all outputs' significance.

output ablation	↓ MAE (std)	↓ MSE (std)	↑ NNSE
full model	**0.190 (0.260)**	**0.104 (0.535)**	**0.556**
no GTSM supervision	0.194 (0.276)	0.114 (0.544)	0.534
GTSM, instead of densification	0.210 (0.246)	0.105 (0.536)	0.554

itime dynamics facilitates short-term forecasting, but that observations further in the past become less informative.

Accuracy vs. lead time To evaluate the effect of the lead time L on performances, we perform inference by varying $L = 4, 6, 8, 10, 12$ hours ahead. Note that the lead time can be systematically varied in a single forward pass thanks to the network being directly conditioned on L. The outcomes in Table 3 validate the intuition that longer lead times exacerbate the prediction problem, affirming its non-linear dynamics and the challenges of medium-term forecasting as encountered by numerical weather prediction models.

Ablation studies We systematically ablate over input information and supervision extent to investigate each element's significance in driving the prediction goodness. The results are reported in Tables 4 and 5, respectively. Input ablations justify the provisioning of coarse GTSM plus ERA5 auxiliary inputs, and show that the network can effectively translate atmospheric reanalysis to ocean states. Furthermore, data dropout is critical for enabling densification and the introduction of flexible lead time conditioning is beneficial. The output ablations confirm that coarse GTSM supervision provides valuable guidance, yet there is additional gains for the densified storm surge output. In sum, the outcomes validate our overall approach as depicted in Fig. 1.

Accuracy vs. sequence length To evaluate the effect of the number of input time points T on performances, we run FiLM U-TAE on input time series of lengths $T = 6, 12, 18, 24$ hours. Table 2 shows that longer sequences drive better forecasts, although gains saturate. This confirms the intuition that more context and data regarding mar-

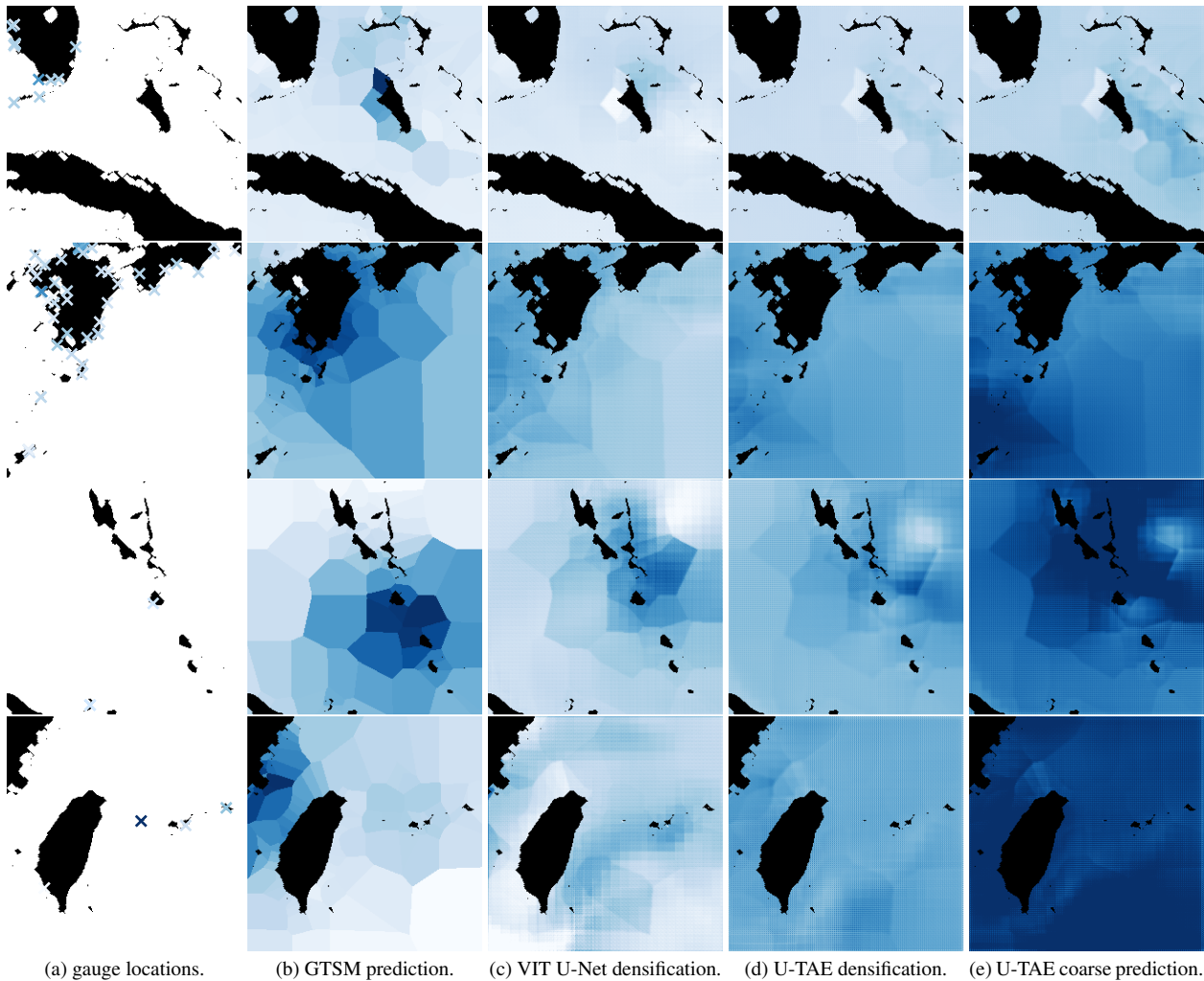| (a) gauge locations. | (b) GTSM prediction. | (c) VIT U-Net densification. | (d) U-TAE densification. | (e) U-TAE coarse prediction. |

Figure 6. **Exemplary data and densified storm surge forecasts** in the densification experimental setup. Rows: Four samples from the test split. Columns: Sampled gauge locations. Dense surge forecasts of GTSM, MaxVit U-Net, FiLM U-TAE and coarse auxiliary FiLM U-TAE predictions. All illustrated outputs are in the densification setup without the target gauge provided, at a lead time of $L = 8$ hours.

Qualitative results Complementary to the reported quantitative outcomes, Fig. 6 illustrates example data and the different densification models' forecasts. Notably, the networks' densified predictions show substantial differences from GTSM, as they are driven by the assimilation of in situ tidal gauge data which GTSM does not incorporate. Specifically, modifications are undertaken close to the shorelines where gauges are present and surge forecasts are most relevant. Furthermore, the densifications often differ from the auxiliary coarse predictions, underlining the functional differences across the two kinds of outputs and highlighting once more the importance of integrating the sparse in situ data. Finally, the appearance of gridding in the spatio-temporal networks' outputs evidences the impact of the ERA5 atmospheric information, which is integrated in the storm surge forecasts.

7. Conclusion

To tackle the aggravating hazard of coastal floods, we introduce a novel dataset and framework forecasting storm surges. Our curated data makes the posed challenge more accessible to the remote sensing community and may serve as a benchmark to fuel future research. Our approach is influenced by recent progress in weather forecasting, and shows that neural networks can implicitly assimilate sparse in situ measurements with coarse weather and ocean state reanalysis products to provide densified forecasts. In a follow-up, we'll explore the operational potential of our approach and replacing retrospective reanalysis products with recently developed forecasting models. Further directions may be the incorporation of satellite altimetry, the modeling of impact at landfall and the translation from storm surges to predicting flood maps. Data and code are given at https://github.com/PatrickESA/StormSurgeCastNet.

References

[1] Marcin Andrychowicz, Lasse Espeholt, Di Li, Samier Merchant, Alex Merose, Fred Zyda, Shreya Agrawal, and Nal Kalchbrenner. Deep learning for day forecasts from sparse observations. *arXiv preprint arXiv:2306.06079*, 2023. 2, 4, 5, 6

[2] Emanuele Bevacqua, Michalis I Vousdoukas, Giuseppe Zappa, Kevin Hodges, Theodore G Shepherd, Douglas Maraun, Lorenzo Mentaschi, and Luc Feyen. More meteorological events that drive compound coastal flooding are projected under climate change. *Communications Earth & Environment*, 1(1):47, 2020. 1

[3] Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970):533–538, 2023. 2, 5

[4] Nicolas Bruneau, Jeff Polton, Joanne Williams, and Jason Holt. Estimation of global coastal sea level extremes using neural networks. *Environmental Research Letters*, 15(7):074030, 2020. 2

[5] Daniel L Codiga. Unified tidal analysis and prediction using the UTide Matlab functions. 2011. 2

[6] Deltares. Delft3D Flexible Mesh Suite. *Delft3D-FLOW, User Manual; Deltares: Delft, The Netherlands*, 2018. 3

[7] Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 4, 6

[8] Avantika Gori, Ning Lin, Dazhi Xi, and Kerry Emanuel. Tropical cyclone climatology change greatly exacerbates US extreme rainfall–surge hazard. *Nature Climate Change*, 12(2):171–178, 2022. 1

[9] Ivan D Haigh, Marta Marcos, Stefan A Talke, Philip L Woodworth, John R Hunter, Ben S Hague, Arne Arns, Elizabeth Bradshaw, and Philip Thompson. GESLA version 3: A major update to the global higher-frequency sea-level dataset. *Geoscience Data Journal*, 10(3):293–314, 2023. 2, 3

[10] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. 1, 2

[11] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3, 6

[12] S. Hoyer and J. Hamman. xarray: N-D labeled arrays and datasets in Python. *Journal of Open Research Software*, 5(1), 2017. 5

[13] Eugenia Kalnay. *Atmospheric modeling, data assimilation and predictability*. Cambridge university press, 2003. 1

[14] Herman WJ Kernkamp, Arthur Van Dam, Guus S Stelling, and Erik D de Goede. Efficient scheme for the shallow water equations on unstructured grids with application to the continental shelf. *Ocean Dynamics*, 61:1175–1188, 2011. 3

[15] Georg Erich Kindermann. Global land / water mask map with 0.3 arc sec. (10 m) resolution. ResearchGate, 2022. Accessed: 2024-01-01. 2

[16] Diederik P Kingma and Jimmy Ba. ADAM: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 6

[17] Ebru Kirezci, Ian R Young, Roshanka Ranasinghe, Sanne Muis, Robert J Nicholls, Daniel Lincke, and Jochen Hinkel. Projections of global-scale extreme sea levels and resulting episodic coastal flooding over the 21st century. *Scientific reports*, 10(1):11629, 2020. 1

[18] Kenneth R Knapp, Michael C Kruk, David H Levinson, Howard J Diamond, and Charles J Neumann. The international best track archive for climate stewardship (IBTrACS) unifying tropical cyclone data. *Bulletin of the American Meteorological Society*, 91(3):363–376, 2010. 3, 5

[19] Frederik Kratzert, Daniel Klotz, Claire Brenner, Karsten Schulz, and Mathew Herrnegger. Rainfall–runoff modelling using long short-term memory (LSTM) networks. *Hydrology and Earth System Sciences*, 22(11):6005–6022, 2018. 3

[20] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023. 2, 5

[21] Gurvan Madec and NEMO team. NEMO ocean engine: Note du pôle de modélisation de l'institut pierre-simon laplace no 27, 2008. 2

[22] Alexandre K Magnan, Michael Oppenheimer, Matthias Garschagen, Maya K Buchanan, Virginie KE Duvat, Donald L Forbes, James D Ford, Erwin Lambert, Jan Petzold, Fabrice G Renaud, et al. Sea level rise risks and societal adaptation benefits in low-lying coastal areas. *Scientific reports*, 12(1):10677, 2022. 1

[23] Sanne Muis, Maialen Apecechea, José Álvarez, Martin Verlaan, Kun Yan, Job Dullaart, Jeroen Aerts, Trang Duong, Rosh Ranasinghe, Dewi Bars, Rein Haarsma, and Malcolm Roberts. Global sea level change time series from 1950 to 2050 derived from reanalysis and high resolution CMIP6 climate projections. In *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*, 2022. 2, 3, 6

[24] J Eamonn Nash and Jonh V Sutcliffe. River flow forecasting through conceptual models: Part I — a discussion of principles. *Journal of Hydrology*, 10(3):282–290, 1970. 6

[25] Grey Nearing, Deborah Cohen, Vusumuzi Dube, Martin Gauch, Oren Gilon, Shaun Harrigan, Avinatan Hassidim, Frederik Kratzert, Asher Metzger, Sella Nevo, et al. AI increases global access to reliable flood forecasts. *arXiv preprint arXiv:2307.16104*, 2023. 2, 3

[26] Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022. 2, 5

[27] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 4, 5, 6

[28] Claudio Persello, Jan Dirk Wegner, Ronny Hänsch, Devis Tuia, Pedram Ghamisi, Mila Koeva, and Gustau Camps-Valls. Deep learning and Earth observation to support the Sustainable Development Goals: Current approaches, open challenges, and future opportunities. *IEEE Geoscience and*

Remote Sensing Magazine, 10(2):172–200, 2022. 1

[29] Russ Rew and Glenn Davis. NetCDF: an interface for scientific data access. *IEEE computer graphics and applications*, 10(4):76–82, 1990. 5

[30] Matthew Rocklin et al. Dask: Parallel computation with blocked algorithms and task scheduling. In *Proceedings of the 14th python in science conference*, volume 130, page 136. SciPy Austin, TX, 2015. 5

[31] Jeffrey D Sachs, Christian Kroll, Guillame Lafortune, Grayson Fuller, and Finn Woelm. *Sustainable Development Report 2022*. Cambridge University Press, 2022. 1

[32] Kaiyue Shan, Yanluan Lin, Pao-Shin Chu, Xiping Yu, and Fengfei Song. Seasonal advance of intense tropical cyclones in a warming climate. *Nature*, 623(7985):83–89, 2023. 3

[33] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015. 3, 6

[34] Casper Kaae Sønderby, Lasse Espeholt, Jonathan Heek, Mostafa Dehghani, Avital Oliver, Tim Salimans, Shreya Agrawal, Jason Hickey, and Nal Kalchbrenner. MetNet: A neural weather model for precipitation forecasting. *arXiv preprint arXiv:2003.12140*, 2020. 2, 3

[35] MG Sotillo, S Cailleau, P Lorente, B Levier, R Aznar, G Reffray, A Amo-Baladrón, J Chanut, M Benkiran, and E Alvarez-Fanjul. The MyOcean IBI ocean forecast and re-analysis systems: operational products and roadmap to the future copernicus service. *Journal of Operational Oceanography*, 8(1):63–79, 2015. 2

[36] Mohsen Taherkhani, Sean Vitousek, Patrick L Barnard, Neil Frazer, Tiffany R Anderson, and Charles H Fletcher. Sea-level rise exponentially increases coastal flood frequency. *Scientific reports*, 10(1):6466, 2020. 1

[37] Timothy Tiggeloven, Anaïs Couasnon, Chiem van Straaten, Sanne Muis, and Philip J Ward. Exploring deep learning capabilities for surge predictions in coastal areas. *Scientific reports*, 11(1):17224, 2021. 2, 3, 6

[38] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *European conference on computer vision*, pages 459–479. Springer, 2022. 4, 6

[39] Jun Wang, Si Yi, Mengya Li, Lei Wang, and Chengcheng Song. Effects of sea level rise, land subsidence, bathymetric change and typhoon tracks on storm flooding in the coastal areas of shanghai. *Science of the total environment*, 621:228–234, 2018. 1

[40] Daniele Zanaga, Ruben Van De Kerchove, Dirk Daems, Wanda De Keersmaecker, Carsten Brockmann, Grit Kirches, Jan Wevers, Oliver Cartus, Maurizio Santoro, Steffen Fritz, et al. ESA WorldCover 10 m 2021 v200. 2022. 2