Correcting Physics-Based Global Tide and Storm Water Level Forecasts with the Temporal Fusion Transformer

A.R. Cerrone[1,2], L.G. Westerink[1,3], G. Ling[1,4], C.P. Blakely[1], D. Wirasaet[1], C. Dawson[2], and J.J. Westerink[1]

[1]Department of Civil & Environmental Engineering & Earth Sciences, University of Notre Dame, Notre Dame, IN, USA 46556
[2]Oden Institute for Computational Engineering & Sciences, University of Texas at Austin, Austin, TX, USA 78712
[3]Department of Computer Science, University of Texas at Austin, Austin, TX, USA 78712
[4]International Research Institute of Disaster Science, Tohoku University, Sendai, Miyagi, Japan

Key Points:

- Global water level forecasting models are subject to several sources of epistemic uncertainty
- Machine-learning-based transformer models can rapidly correct for model discrepancy stemming from systematic bias and epistemic uncertainty
- Coupling transformer models to high-fidelity global water-level models is highly effective when supplemented with physics-based covariates

Corresponding author: Albert Cerrone, `acerrone@nd.edu`

Abstract

Global and coastal ocean surface water elevation prediction skill has advanced considerably with improved algorithms, more refined discretizations and high-performance parallel computing. Model skill is tied to mesh resolution, the accuracy of specified bathymetry/topography, dissipation parameterizations, air-sea drag formulations, and the fidelity of forcing functions. Wind forcing skill can be particularly prone to errors, especially at the land-ocean interface. The resulting biases and errors can be addressed holistically with a machine-learning (ML) approach. Herein, we weakly couple the Temporal Fusion Transformer to the National Oceanic and Atmospheric Administration's (NOAA) Storm and Tide Operational Forecast System (STOFS 2D Global) to improve its forecasting skill throughout a 7-day horizon. We demonstrate the transformer's ability to enrich the hydrodynamic model's output at 228 observed water level stations operated by NOAA's National Ocean Service. We conclude that the transformer is a rapid way to correct STOFS 2D Global forecasted water levels provided that sufficient covariates are supplied. For stations in wind-dominant areas, we demonstrate that including past and future wind-speed covariates make for a more skillful forecast. In general, while the transformer renders consistent corrections at both tidally and wind-dominant stations, it does so most aggressively at tidally-dominant stations. We show notable improvements in Alaska and the Atlantic and Pacific seaboards of the United States. We evaluate several transformers instantiated with different hyperparameters, covariates, and training data to provide guidance on how to enhance performance.

Plain Language Summary

Forecasted water levels in coastal regions are used to predict flooding risk, currents that impact navigation, the transport of nutrients and pollutants, and the conditions that make coastal waters favorable for fisheries. Computational models are typically exercised to render water-level forecasts as far as seven days into the future. While the physics underlying these models is relatively well understood, the data flowing into the models can be subject to uncertainty. For example, water depth is a sensitive model parameter, yet even the best LiDar-derived data sets may have significant errors. Model and meteorological wind speed data is another parameter prone to bias and uncertainty. In some cases, these models lack the ability to resolve oceanic and hydrological processes, phenomena that contribute to coastal water levels. Herein, we propose a fairly simple machine-learning-based strategy to improve the predictive capacity of these models in the presence of these limitations. In particular, we couple the ocean hydrodynamic model ADCIRC to a transformer to improve ADCIRC's predictions. We show that the two models individually have larger error bands than when the are combined.

1 Introduction

Water levels in the coastal ocean and its adjacent floodplain impact navigation, water quality, fisheries, and livelihoods in coastal communities. Water motion is driven by highly predictable gravitational forces of the moon and sun acting on the earth's ocean water; by more chaotic forces induced by wind, atmospheric pressure, wind waves, rainfall and regional hydrology; and forces derived from the ocean's thermohaline structure with corresponding ocean current systems and ice packs. The governing physics was pioneered by Laplace (1776) and later expanded by Barré de Saint Venant and Boussinesq (Hager et al., 2019) leading to the depth integrated shallow water equations (Hervouet, 2007). More detailed three dimensional forms of the conservation laws were derived from the Navier-Stokes equations, and are typically subject to the assumption of a hydrostatic pressure approximation and the Boussinesq approximation for density (Roelvink & van Banning, 1995; Haidvogel et al., 2000; Hervouet, 2007). Given sufficient resolution of the geometric and bathymetric intricacies of the coastal ocean and especially of the adjacent

floodplain with its narrow tidal inlets, estuaries, and dendritic channel networks and rivers, the two dimensional shallow water equations generally simulate coastal water levels remarkably well. This has especially been the case as open water boundary condition specification uncertainty has been eliminated by evolving from regional to global domain models, as was initially envisioned by Laplace. Additionally, closure relationships focus on parameterizing bottom boundary layer dissipation, internal tide generation and subsequent dissipation, viscous turbulent dissipation, and air-sea and air-ice-sea momentum transfer, which are all largely empirically derived from observed data.

Global total water level models include Deltares' Delft3D based Global Tide and Surge Model (GTSM) (Verlaan et al., 2015; De Kleermaeker et al., 2017), Environment and Climate Change Canada's NEMO-based model (Wang et al., 2021; Wang & Bernier, 2023), and NOAA's ADCIRC-based unstructured mesh Storm and Tide Operational Forecast System 2D Global (STOFS 2D Global) model (Seroka et al., 2023). STOFS 2D Global, the model applied in this study, is a barotropic implementation of the shallow water equations with resolution ranging from 80 m within intricate inlets to 25 km across the abyssal plains of the deep ocean (Pringle et al., 2021; Blakely et al., 2022). The model leads to excellent predictability of tides with the global observed-to-modeled M_2 tide compared at 236 deep water stations resulting in a coefficient of determination, R^2, equal to .985, a mean amplitude error of 2.4 cm, and a normalized root mean square error (NRMSE) equal to 0.075. At 449 shelf stations, the model yields an R^2 equal to .984, a mean amplitude error of 4.3 cm, and a NRMSE equal to 0.084. STOFS 2D Global is to date the most accurate published non-data assimilated model with respect to tides (Stammer et al., 2014; Blakely et al., 2022). A multiyear hindcast (see Section 2) indicates that surface water elevation can be predicted at 213 U.S. NOAA National Ocean Service (NOS) water level stations with an R^2 equal to 0.94, an average absolute error equal to 7.3 cm, and a NRMSE equal to 0.21.

Model accuracy is influenced by factors such as geometric representation and mesh resolution, topo-bathymetry, and specified values for the parameterized dissipation terms. Additionally, the underlying physical processes that are incorporated into STOFS 2D Global are subject to formulations errors. Some salient processes may be missing altogether. The biggest source of epistemic uncertainty associated with STOFS 2D Global is its meteorological forcing. Available wind fields, which are essential to forecasting accurate water levels, are typically spatially and temporally coarse, often leading to significant wind and associated water level prediction errors over shallow inland waterbodies. Moreover, model winds tend to be muted for intense tropical cyclones which result in muted storm surge forecasts. Collectively, these varied sources of model discrepancy motivate an improvement capability that scrutinizes simultaneously STOFS 2D Global output, its input (e.g. forcings), and past observations.

Generally, Bayesian methods have been adopted to improve model accuracy amidst quantifiable uncertainty. Typically, for spatially and/or temporally invariant applications, algorithms like the Extended Kalman filter (Holland, 2020) and Particle filter (Ristic et al., 2003; Li et al., 2017; Elfring et al., 2021) have been leveraged to render these improvements. While they have been adopted for spatiotemporal domains (Butler et al., 2012, 2015; Rougier et al., 2023) and will invariably be scaled for operations in the future, due to computational complexity and expense, they cannot be deployed today for rapid hydrodynamic model improvement. We propose to carry out model improvement by means of machine learning with a transformer model. Rather than operating on the model directly, as is commonly done in a Bayesian framework, the transformer would enrich the model outputs themselves (see schematic diagrams in Figure 1). By moving to this machine-learning-based approach, although we lose some accuracy that would come from a tight coupling required by a traditional Bayesian scheme, we gain processing speed. We develop and demonstrate this approach in computing corrected water levels for the STOFS 2D Global model.

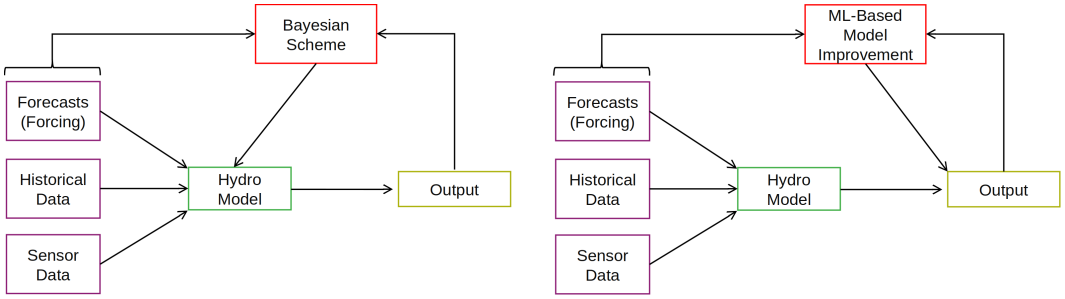

Figure 1. Process maps incorporating a Bayesian scheme (left) and ML (right) for model improvement.

The machine learning (ML) model considered herein is the temporal fusion transformer (TFT) (Lim et al., 2021). The TFT is a transformer-based model (Vaswani et al., 2017) that uses self-attention to learn long-term dependencies. The transformer is a compelling option for hydrodynamic applications because some coastal regions experience predictable, yet "longer term" water elevation patterns that recurrent neural networks (RNNs) or long short term memory (LSTM) cells alone could not otherwise capture. Granted, in the present application, the transformer would be compelled to scrutinize patterns in model error space, but like the surface water elevations themselves, these error signals are also expected to be repeated. Moreover, the TFT can accommodate static, past, and future covariates. These covariates enable the TFT to consider simultaneously the relationship between observed and predicted water levels and other relevant quantities including wind speed and location. In essence, referring back to Figure 1, these covariates are the data proceeding from model inputs and outputs. Consequently, as the transformer scrutinizes patterns in model error space, it also attends to model inputs and outputs, drawing parallels to a traditional Bayesian scheme.

ML models have featured quite extensively in coastal water level modeling. For example, in an early application, de Oliveira et al. (2009) considered a multilayer perceptron (MLP) to predict storm surge at a single station in Southeast Brazil and demonstrated reasonable performance out to 24-hour forecast horizons. Later, Ayyad et al. (2022) assessed seven ML models to predict peak storm surge height caused by tropical cyclones in the New York Metropolitan Area. They trained their models against output from the model considered in this study, ADCIRC. They determined that of the seven ML models, a support vector regressor (SVR) and an ensemble of decision trees with adaptive boosting were the most performant. Xie et al. (2023) used a convolutional neural network (CNN) to ingest two-dimensional wind forcing for single-site water level elevation forecasting in Southeast China. They demonstrated strong performance out to three-day forecasting horizons provided that a 24-hour water-level input is supplied. Tiggeloven et al. (2021) considered an ensemble of different models which included the coupled CNN and LSTM (ConvLSTM) to predict surge levels globally at more than 700 tidal stations. They also provided meteorological data as covariates. They demonstrated that the LSTM outperformed the other models, but that the CNN, provided that it was instantiated with a sufficient number of hidden layers, had the potential to outperform the lot despite its increased compute time. Additionally, they showed that their models generally performed better for higher-latitude tidal stations. Most recently, Pachev et al. (2023) demonstrated a two-tiered location-agnostic approach to predict peak storm surge from tropical cyclones. First, they exercised a classifier to identify inundated points. Thereafter, they ran a neural network (trained with boosting) to predict the level of inundation.

In this paper, we deviate our approach from these coastal water level studies. While we predict observed water levels, we do so by weakly coupling a numerical model (viz.

STOFS 2D Global solved by ADCIRC) to the TFT. We show that compared to ADCIRC alone, this coupling results in improved station-based water level forecasts out to 7-days. This type of coupling is not without precedent, especially in atmospheric modeling. For example, Bonavita and Laloyaux (2020) demonstrated that a MLP can extend the capability of the weak-constraint formulation of the 4D-Var data assimilation framework to the troposphere. Zampieri et al. (2023) used a fully-connected MLP trained on observations to reduce a temperature bias over Arctic sea ice in reanalysis products. In ocean modeling, Bolton and Zanna (2019) trained CNNs on output from a quasi-geostrophic ocean model to predict eddy momentum forcing and determined that training on turbulent regions leads to enhanced CNN extendability. More relevant to the present discussion, they conjecture that CNNs can be coupled to sparse interpolated observational data to render accurate predictions of large-scale flow in the presence of turbulence.

In this study, we train the TFT to correct surface water elevation, η, predictions made by ADCIRC-based STOFS 2D Global at 228 NOAA water level stations in the northern half of the Western Hemisphere. We use a three-year STOFS 2D Global hindcast and 6-minute resolution NOAA water level observations to instantiate the TFTs considered herein. In Section 2, we summarize STOFS 2D Global and the aforementioned three-year hindcast. Additionally, we detail the ML framework considered herein including model architecture, data processing, and training. Thereafter, in Section 3, we detail its performance and probe its sensitivity to training set size, various hyperparameters, covariates, and regionality. Finally, in Section 4, we discuss how this framework could be extended to accommodate extreme weather events and be adapted for both probabilistic and off-station STOFS 2D Global correction.

2 Materials and Methods

2.1 STOFS 2D Global

2.1.1 Overview

STOFS 2D Global, NOAA's operational global water level model, is based on an optimized unstructured mesh of the global ocean with the highest resolution focused on U.S. coastal waters and floodplains. The model applies 25-km resolution across deep ocean abyssal plains, down to 2.5-km resolution across steep ocean topography, and refines coasts down to 2.5 km globally (Blakely et al., 2022). U.S. inland waters and coastal features are resolved down to 80 to 125 m. This includes a large extent of the floodplains, coastal estuaries, inland channels, and levee systems of the contiguous United States, Puerto Rico, Alaska, Hawaii, and Micronesia. Figures 2 and 3 depict mesh resolution and bathymetry in the global model. It is noteworthy that smaller elements discretize mid-ocean ridges, shelf breaks, and submerged mountain chains in order to improve internal tide dissipation models (Pringle et al., 2021; Blakely et al., 2022). Global and regional bathymetric datasets were applied including GEBCO2020 (IHO-UNESCO, 2020), RTopo-2 (Schaffer et al., 2016), Canadian CHS-NONNA100 (Fisheries & Canada, 2023), nthaus100 Grid (Beaman, 2018), and the Allen Coral Atlas (Lyons et al., 2022). In US water, various regional bathymetric data sources were used from the United States Geological Survey (USGS), NOAA, U.S. Army Corps of Engineers, and Northeast Ocean Data. The melded bathymetry is mesh scale averaged at nodes, Figure 3.

The modeling system is forced with tidal potential functions, associated self attraction and load terms (via FES2014), winds, atmospheric pressure, and sea ice from NOAA's Global Forecast System (NOAA-EMC, 2023). Friction force parameterizations for bottom boundary layer dissipation, internal tide generation/dissipation, air-sea wind drag, and air-ice-sea drag are based on previous regional studies and were optimized for STOFS 2D Global (Dietrich et al., 2011; Chen et al., 2013; Hope et al., 2013; Kerr, Donahue, et al., 2013; Kerr, Martyr, et al., 2013; Pringle et al., 2018, 2021; Blakely et al., 2022).

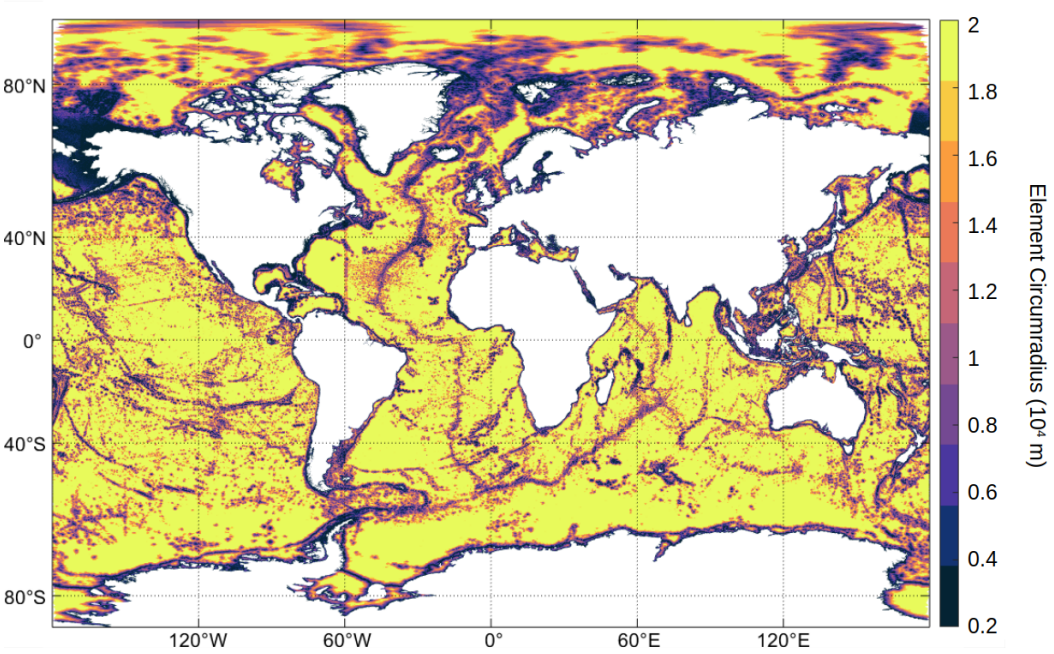

Figure 2. STOFS 2D Global mesh resolution. This mesh consists of 12,784,991 nodes and 24,875,313 elements.

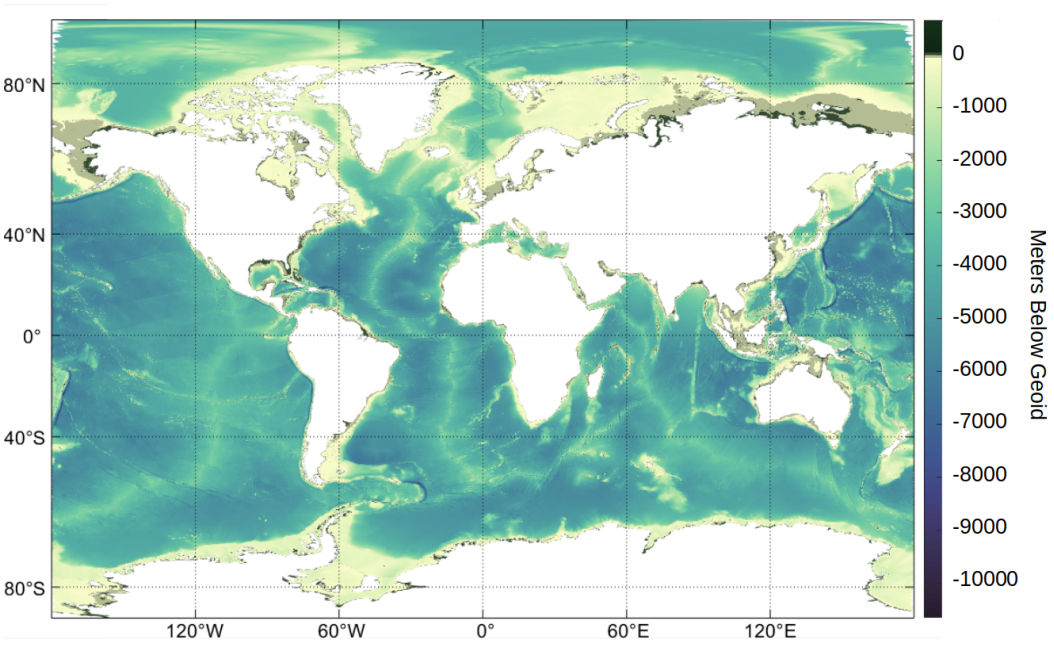

Figure 3. STOFS 2D Global bathymetry melded from various global and regional sources.

NOAA's National Centers for Environmental Prediction (NCEP) and National Ocean Service (NOS) Coast Survey and Development Laboratory (CSDL) jointly operate the model to produce four-times daily 180-hour water level forecasts (NOAA-OPC, 2023; NOAA-NOS, 2023). The University of Notre Dame Computational Hydraulics Laboratory also runs a "shadow" model 168-hour once-daily forecast (Wood, 2023) to compare to ongoing development versions. The model has undergone a suite of improvements, for example, by including more levee systems, increasing spatial resolution, refining bottom friction parameterizations, and increasing both temporal and spatial resolution of atmospheric forcing (Pe'eri, 2023).

STOFS 2D Global is driven by ADCIRC, a community finite-element 2D and 3D hydrodynamics solver (*ADCIRC*, 2023) used for modeling tides and coastal storm surge flooding at local and regional scales (Westerink et al., 2008; Bunya et al., 2010; Hope et al., 2013). STOFS 2D Global applies ADCIRC's two-dimensional barotropic solver. To facilitate solutions on the global domain, the model was updated to include a generalized cylindrical mapping system for transforming spherical coordinates to a rectilinear system. Since the spherical coordinate system has a singularity at the poles, and the cylindrical mapping system likewise does not permit elements spanning over the poles, a coordinate rotation is added that places both poles overland and likewise rotates the Coriolis, surface wind, and internal wave drag terms (Pringle et al., 2021).

The operational implementation is not currently forced with hydrology or with the ocean's thermohaline circulation, both of which can impact water levels on longer-term periods (Pringle et al., 2018). To accommodate these slower time-scale fluctuations, the modeling system computes the mean water level for the five days prior to the start of any forecast for both the model and the NOS station measurement, and levels the forecast water levels accordingly. Fast time scale fluctuations associated with hydrology (e.g. a high intensity local rainfall event in a small scale channel) or with the ocean's thermohaline system (e.g. the changes in the vertical density structure caused by a passing hurricane) will appear as additional errors in the forecast. Furthermore, the global model is not presently coupled to wind wave models and therefore is not forced with wave radiation induced stresses associated with wind wave transformation and breaking. The wind wave model coupling is typically incorporated into regional models and boosts water levels along coasts by between 5 cm and up to 0.5 m in limited regions (Dietrich et al., 2011; Hope et al., 2013; Joyce, Gonzalez-Lopez, et al., 2019; Joyce, Pringle, et al., 2019). Again, this will appear as a missing physics bias in the model, although typically this bias will be highly correlated to the specific coastal geometry and bathymetry and wind intensity and direction. We apply the five-day prior water level adjustment to the seven day forecast horizon in this study.

2.1.2 Three-Year Hindcast

A three-year STOFS 2D Global hindcast was run for the period of September 2016 - September 2019 with one month spin-up and a 6-second time step. As discussed in Section 3, the hindcast was used to train and validate the ML framework. We exercised ADCIRC version 55. The internal tide dissipation and boundary layer dissipation parameters were adopted from Blakely et al. (2022). The following options were indicated: ICS=-22 (Mercator projection with pole rotation), IM=511113 (implicit mode), A00=0.8, B00=0.2, C00=0.0 (time weighting factors), H0=0.1 (minimum water depth), TAU0=0.05 (Generalized Wave-Continuity Equation, GWCE, weighting factor that weights the relative contribution of the primitive and wave portions of the GWCE), NTIP=2 (tidal potential and self attraction / load tide forcings are used), and DT=6 (simulation time step in seconds). Eight dominant astronomical tidal harmonic constituents (M_2, N_2, S_2, K_2, K_1, Q_1, O_1, P_1) were forced using the tidal potential function as well as self-attraction and loading. Additionally, atmospheric forcing was sourced from the NCEP Coupled Fore-

cast System Model version 2 (CFSv2), a reanalysis product with 0.25-deg spatial resolution.

The hindcast was run on the Texas Advanced Computing Center's Frontera supercomputer (Intel Xeon Platinum 8280, clock rate 2.7Ghz, Peak node performance 4.8TF, double precision).

2.2 ML Framework

2.2.1 General TFT Operation

The TFT models exercised herein were based on the formulation given by Lim et al. (2021). We used the Python library Darts (Herzen et al., 2022) to train and evaluate them. The TFT architecture possesses multiple desirable components for spatiotemporal forecasting tasks. TFTs employ variable selection networks to ensure that relevant input is considered at a given time step. Meanwhile, LSTM layers learn short-term temporal dependencies and allow for heterogeneous data sources to be encoded as static and dynamic covariates. Gating mechanisms let the model ignore architecture components where appropriate, while multi-head attention layers provide the model the ability to learn interpretable long-range time dependencies. While the original TFT implementation uses quantile regression to output probabilistic forecasts, we primarily used a loss function in place of likelihood to make deterministic predictions. The hidden layer size, which is shared across both LSTM and self-attention layers, the number of attention heads, the number of LSTM layers, and the learning rate for our model were tuned in a hyperparameteric sweep discussed later in this section.

The target of each TFT was the signed difference between the observed surface water elevation (from NOAA observed station time histories) and the predicted surface water elevation (from the three-year hindcast). Each TFT forecasted (i.e. decoded) this target seven days (168 hours) into the future with 1-hour temporal resolution. It made this forecast in a "single shot" as opposed to autoregressively. We considered a 120-length (120 hour = 5 days) input or encoding region. To facilitate each TFT making this forecast, we supplied user-defined covariates. We supplied them for a period prior to the forecasting horizon (i.e. past covariates), within the forecasting horizon (i.e. future covariates), and also invariantly of time (i.e. static covariates), Table 1. It is noteworthy that three time covariates (hour of the day, day of the week, and month of the year) were included to facilitate the model learning seasonal, diurnal, and semi-diurnal trends. The static covariates, in turn, were provided to suggest spatial coherence. The physics-based covariates, finally, were considered to facilitate predictions at wind-dominant stations. In a post-processing step extrinsic to each TFT, the TFT's output was added to ADCIRC's raw prediction to render a "ML-Corrected" ADCIRC prediction.

It is noteworthy that the hindcast-centric treatment of the TFT provided herein deviates slightly from how this methodology could be applied operationally for forecasting, Figure 4. Here, we used hindcasted ADCIRC predicted water levels and hindcasted meteorology in place of the forecasted future covariates. Leveraging these hindcasted covariates enabled us to bypass the late-horizon uncertainty stemming from our meteorological forcing. Consequently, the results provided herein indicate intrinsic TFT performance unblemished by spurious meteorological forcing; however, in forecasting mode, this uncertainty would likely lead to degraded late-horizon performance.

2.2.2 Data Processing

The study considered surface water elevations from a three-year ADCIRC-driven STOFS 2D Global hindcast (Section 2.2) and observed surface water elevations at NOAA stations. Both products maintain a 6-minute temporal resolution. The 228 stations, plotted and characterized with regard to tidal or wind dominance in Figure 5, were distributed

Table 1. Past, future, and static covariates considered by the TFT model.

Covariate	Type	Kind	Description
Hour of the Day	Past & Future	Time	Integer from 1-24
Day of the Week	Past & Future	Time	Integer from 1-7
Month of the Year	Past & Future	Time	Integer from 1-12
10-m U Wind Speed (CFSv2)	Past & Future	Dynamic Physics	Float
10-m V Wind Speed (CFSv2)	Past & Future	Dynamic Physics	Float
Tide Resynthesis	Past & Future	Dynamic Physics	Float
η Prediction (ADCIRC)	Past & Future	Dynamic Physics	Float
Latitude	Static	Location	Float
Longitude	Static	Location	Float

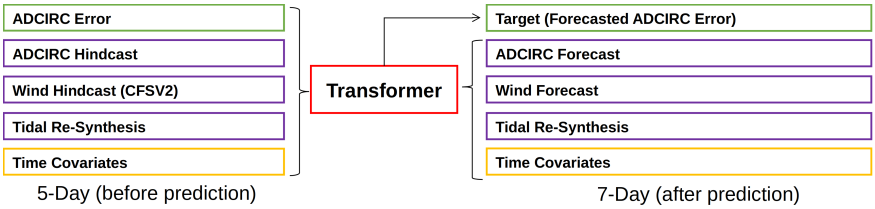

Figure 4. Diagram of transformer in an operational forecasting scheme.

in four regions: (1) Alaska, (2) West: US Pacific Seaboard, Hawaii, Midway Atoll, (3) Gulf: Gulf of Mexico, and (4) East: US Atlantic Seaboard, Puerto Rico, Bermuda. Most of the stations are tidally-dominant (i.e. the tidal potential energy was on the order of the total potential energy at these locations); however, some are situated in strongly wind-dominated locations (i.e. shallow locations wherein water levels are sensitive to the prevailing winds). The Gulf region is composed almost exclusively of wind-dominant stations while the East region includes several in the shallow Chesapeake and Delaware Bays. When we evaluate performance in Section 3, we will discriminate tidally-dominant from wind-dominant stations.

We mapped meteorological forcing (from CFSv2) and model surface water elevation predictions to these stations during the hindcast simulation. These data were then extracted at the top of each hour. The corresponding NOAA observed surface water elevations at the top of each hour were then matched to these data. Model and observed water levels were with respect to the mean sea level (MSL) vertical datum. Due to missing NOAA observed levels for some stations, this matching was incomplete. Given that we did not fill in missing data, the entirety of the hindcast period for some stations could not be used for training / testing. Rather, we segregated the matched data into time-contiguous chunks.

We trained each TFT on a series of time-contiguous chunks. The chunking was accomplished by first segregating the matched data into time-contiguous spans that were at least longer than the desired chunk size (i.e. number of hours). For example, the 5-day encoding region and 7-day decoding region considered herein constituted a chunk size of 288. Thereafter, for each span, a window with length equal to the chunk size traversed the span incrementally. For each increment, the chunk within the window was extracted either for training, validation, or testing. It moved chronologically along the span with a predefined shift length. A shift length equal to the chunk size would result in no chunk overlap; a shift length less than the chunk size would result in chunk overlap and more extracted chunks. We adopted a 1-day shift length to guarantee some overlap.

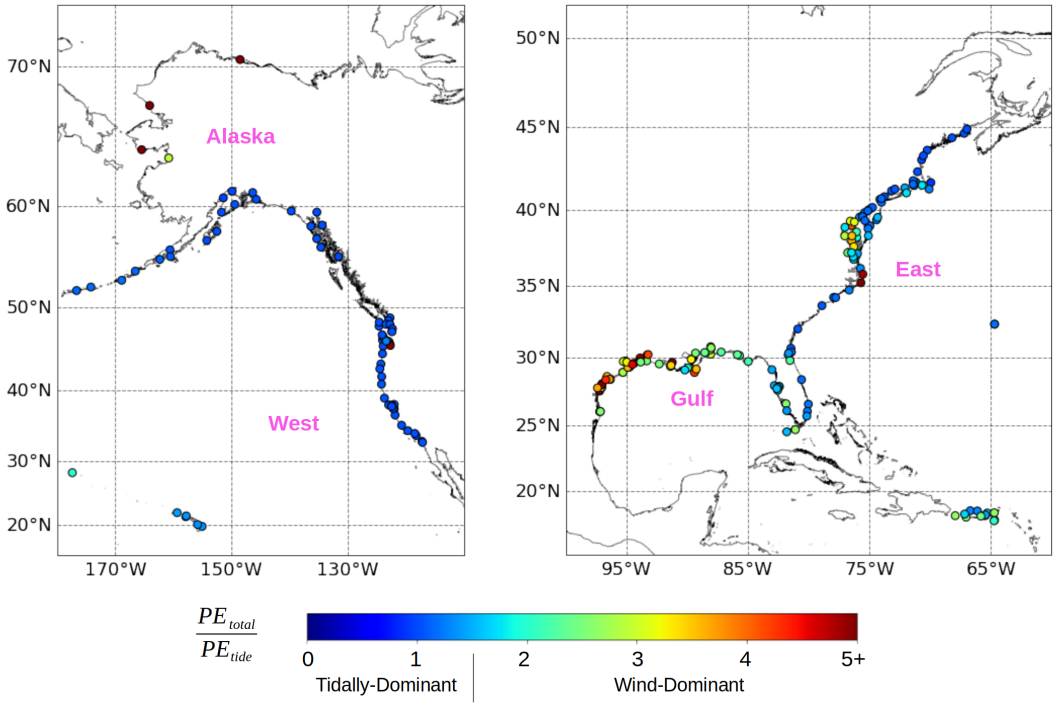

Figure 5. NOAA observed surface water elevation stations considered in study. Stations were grouped into the following regions to facilitate evaluation: (1) Alaska, (2) the US Pacific Seaboard, Hawaii, and Midway Atoll denoted "West", (3) the Gulf of Mexico denoted "Gulf", and (4) the US Atlantic Seaboard, Puerto Rico, and Bermuda denoted "East". Colors reference the ratio between a given station's total potential energy to its tidal potential energy. Values close to 1 indicate tidal dominance. Higher values suggest wind-dominance.

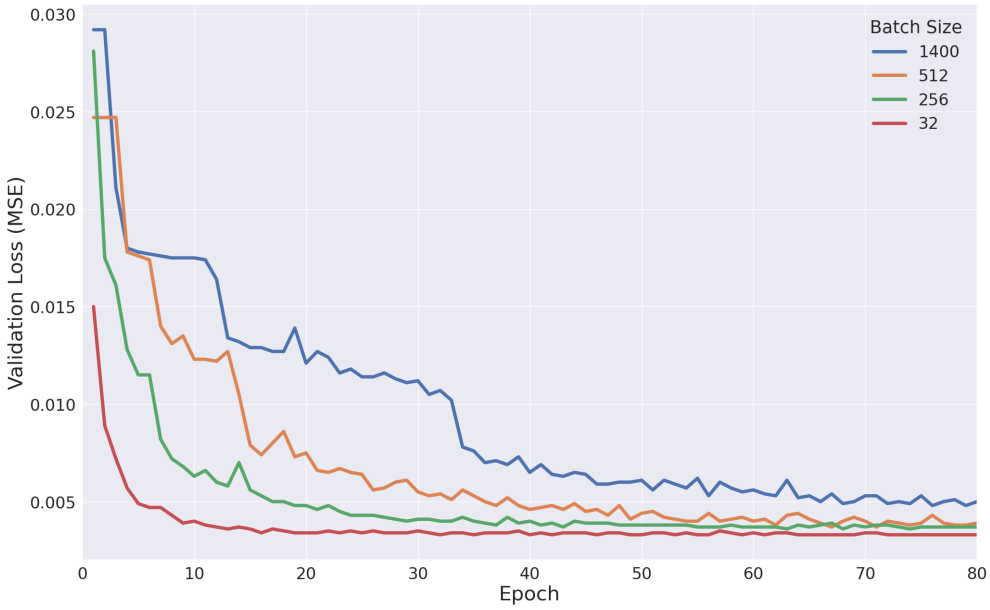

Figure 6. Validation loss plotted as a function of epoch for various batch sizes. Here, validation loss is the average mean squared error over the validation set. Only stations in the Northeast are considered.

2.2.3 TFT Training and Hyperparametric Sweep

Unless otherwise noted, for each station, we selected the first 70% of the chunks in the hindcast period for training. The remaining chunks were randomly shuffled into validation and test sets (20% validation, 10% testing). This ensured that all of the stations had representation in both training, validation, and testing; however, depending on NOAA data coverage, some stations had more training / validation / testing chunks than others. We trained each TFT on all of the designated training chunks station-by-station in chronological order.

During preliminary studies, we found TFT performance to be largely invariant to batch size (i.e. number of chunks passed to GPU for forward and back propagation at a time), but smaller batch sizes resulted in slightly enhanced performance over a reduced number of epochs (i.e. number of full passes through the training set), Figure 6. We hypothesize that the greater number of gradient updates from smaller batch sizes facilitated optimization. However, smaller batches on performant GPUs required longer training duration. Consequently, we settled on intermediate batch sizes as they yielded similar overall results with reduced wall-clock time. To facilitate the optimizer's back propagation, we ensured that each batch was composed primarily of chunks from the same region (Figure 5). For example, if the majority of the chunks in a batch were from tidally-dominant stations, the few from less-predictable wind-dominant stations would likely not contribute to that batch's loss in any significant manner. Batching with same-region chunks was conducted to mitigate this issue. We also interrogated several optimizers including stochastic gradient descent, Adadelta, Adagrad, and quasi-Newton methods, but Adam proved to be the most stable and performant.

We adopted the mean squared error (MSE) loss function in place of mean absolute error (MAE) loss to aggressively minimize outliers. In exploratory studies, we observed that both MAE and MSE loss targeted intermediate errors equally effectively, but that MSE loss reduced the larger errors at some wind-dominant stations (e.g. Annapo-

lis, Baltimore) more reliably. We could have trained the TFTs with quantile regression to produce probabilistic forecasts; however, given that our focus herein was rendering deterministic predictions, we did not pursue this option.

Upon settling on a batching strategy, optimizer, and loss function, automated Bayesian optimization was employed via a Tree-Structured Parzen Estimator (TPE) to efficiently minimize MSE loss over the following hyperparameters: number of LSTM layers, number of attention heads, hidden layer dimension, and dropout rate. Our model proved relatively insensitive to these hyperparameters over a subset of stations in the Northeast; however, modest performance gains were obtained nonetheless. Based on this exercise, three LSTM layers, three attention heads, and a dropout rate of approximately 10% were adopted. Hidden layer size was ultimately capped at 110 to keep training time and computational costs manageable.

After conducting the hyperparametric sweep, ten models were identified for comprehensive training and evaluation. They are described below:

1. Global - All - Baseline: This model inherited the tuned hyperparameters in addition to all of the covariates listed in Table 1 (U, V, ADCIRC η, tidal resynthesis and time and spatial covariates). It was trained on the first 70% of each station's chunks in the hindcast period. It was trained on all stations in the Alaska, West, Gulf, and East regions. In the model, the encoder attends to both past and future covariates, while the decoder attends only to future covariates.
2. Global - No Physics: This model was identical to the Baseline model incorporating the time and spatial covariates, but did not retain any of the physics-based dynamic covariates (U, V, ADCIRC η, tidal resynthesis).
3. Global - No Tides: This model was identical to the Baseline model, but was not given access to the tidal covariates. Using this model, we wanted to assess if tidal-centric errors could be addressed by the transformer in the absence of a tidal resynthesis.
4. Global - No Winds: This model was identical to the Baseline model, but lacked past and future wind covariates.
5. Global - All - Full Attention: This model derived from the Baseline model, but adopted a slightly different architecture wherein the decoder was allowed to attend to previous, current, and forthcoming future covariates in the forecast horizon. Enabling the decoder to attend to current and forthcoming future wind data, and not simply previous covariates as is done in the baseline model, was hypothesized to increase performance in wind-dominant stations.
6. Global - All - 30% Train: This model derived from the Baseline model, but was trained on only the first 30% of each station's chunks in the hindcast period (as opposed to 70%).
7. Alaska - All: This model derived from the Baseline model, but was trained exclusively on chunks in the Alaska region.
8. West - All: This model derived from the Baseline model, but was trained exclusively on chunks in the West region.
9. Gulf - All: This model derived from the Baseline model, but was trained exclusively on chunks in the Gulf region.
10. East - All: This model derived from the Baseline model, but was trained exclusively on chunks in the East region.

We leveraged Nvidia A6000 GPUs for training and inference. Training a "global" model required approximately 24-hours of wall-clock time. Evaluating a "global" model's entire test set for all chunks therein required only a minute of wall-clock time.

3 Results

3.1 Evaluation Metrics

To quantify the performance of each TFT model, we considered four evaluation metrics: normalized root-mean-squared error (NRMSE), maximum error (MAX), coefficient of determination (R^2), and the Willmott skill score (WSS) (Willmott, 1981). They are given by:

$$NRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\eta_{obs,i} - \eta_{pred,i})^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\eta_{obs,i})^2}} \tag{1}$$

$$MAX = max(|\eta_{obs,i} - \eta_{pred,i}|) \tag{2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\eta_{obs,i} - \eta_{pred,i})^2}{\sum_{i=1}^{n}(\eta_{obs,i} - \eta_{obs,avg})^2} \tag{3}$$

$$WSS = 1 - \frac{\sum_{i=1}^{n}(\eta_{obs,i} - \eta_{pred,i})^2}{\sum_{i=1}^{n}(|\eta_{pred,i} - \eta_{obs,avg}| + |\eta_{obs,i} - \eta_{obs,avg}|)^2} \tag{4}$$

Note that η_{obs} and η_{pred} denote observed and predicted surface water elevations, respectively, and n denotes output chunk length (168). We consider two sets of predicted water levels: ADCIRC-predicted (i.e. raw ADCIRC) and ML-Corrected (i.e. ADCIRC+ML). NRMSE and MAX are error metrics. Smaller values denote superior model performance. R^2 and WSS are regressive score metrics. For both, a value of 1 denotes perfect skill. These four metrics were used to evaluate individual chunks.

3.2 Evaluated Performance

In this section, we evaluate TFT region-based and station-based performance using the test set of each station. This set was separate from training and validation. Moreover, for each station, the test set period did not overlap with that of the training set. On average, each station had 100 test chunks. Each evaluation considered evaluation metrics calculated for individual chunks, and then these metrics were averaged over regions (see region-based performance) or individual stations (see station-based performance). Figure 7 illustrates how ADIRC + ML improves the solution as compared to ADCIRC alone for one 7-day forecasting chunk at three sample stations.

3.2.1 Region-Based Performance

We calculated the four evaluation metrics for the 7-day forecast horizon in each of the chunks in our test set. Thereafter, we averaged these metrics over all of the chunks at each station. We then grouped these averaged metrics by region. The resulting distributions are plotted in Figure 8. It is clear that all of the TFTs rendered relatively aggressive corrections in Alaska, the West (predominantly Pacific Seaboard), and the East (predominantly Atlantic Seaboard). This is likely because each of these regions have a majority of tidally-dominant stations wherein surface water elevations are determined by highly deterministic, cyclical tidal potential functions and not by lower fidelity, less certain meteorology. The performance of the TFT devoid of physics-based dynamic covariates supports this hypothesis as it rendered relatively aggressive corrections correlating errors predominantly to time covariates.

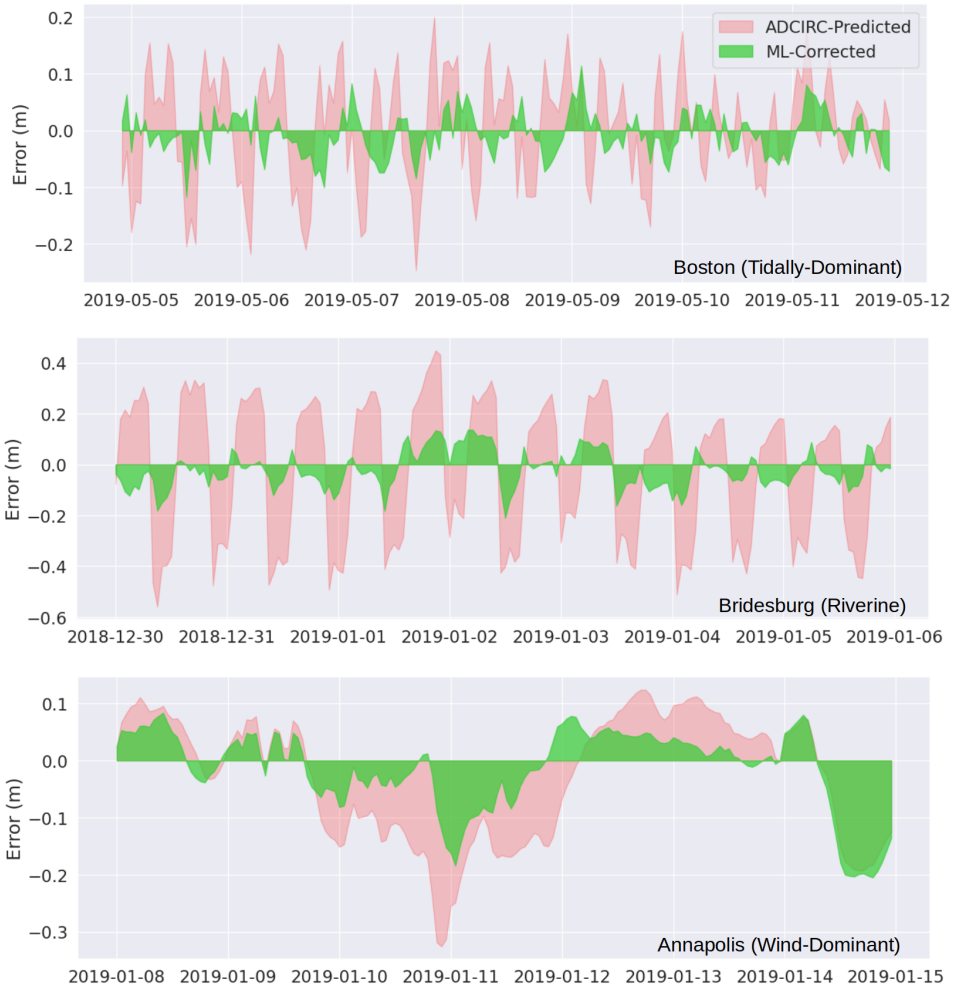

Figure 7. Examples of 7-day output test chunks considered in the evaluation. ADCIRC-predicted and ML-corrected error are considered here. Values closer to x-axis represent higher skill. ML output was produced in a "single shot" for each chunk.

In the Gulf of Mexico, however, the TFT-based corrections were less aggressive. The Gulf has a majority of wind-dominant stations which are difficult to predict, leading to larger ADCIRC prediction uncertainty and weaker ML-based corrections. Adding tides and ADCIRC's predictions ("Global - No Winds") and thereafter adding wind ("Global - All - Baseline") clearly made for more performant TFTs, especially in the Gulf. The TFT trained with full attention demonstrated roughly the same skill as its counterpart with limited attention, suggesting that allowing the decoder to attend to current and forthcoming future covariates does not enhance performance and that the encoder's attention to past and future covariates controls model skill. Excluding resynthesized tides but retaining all other covariates ("Global - No Tides") rendered similar performance to the other TFTs supplied with physics-based covariates, suggesting that time covariates are facilitating tidal pattern recognition in ADCIRC's error space and surface water elevation prediction. The TFT with a reduced training set size exhibited approximately the same skill as the TFT devoid of physics-based dynamic covariates. Finally, the regional models demonstrated slightly worse performance compared to the "Baseline" model.

From this high-level region-centric evaluation, we can assume that (1) the TFT is intrinsically capable of correcting tidally dominant stations without physics-based dynamic covariates, but by adding additional physics-based covariates (viz. ADCIRC prediction), the TFT renders a more aggressive correction and (2) the physics-based covariates facilitate improved TFT performance at wind-dominated stations. We will confirm these assumptions in the following section wherein we conduct a station-based evaluation.

3.2.2 Station-Based Performance

The evaluation metrics were calculated for each station's chunks in our test set. First, we considered TFT late-horizon performance. All time-series forecasting models suffer from degradation, and this is typically correlated to horizon length. The TFT is no exception. As shown in Figure 9, TFT skill within the 6-7 day horizon was lower than its skill within the 0-1 day horizon. Moreover, more prominent drops in late-horizon skill generally occurred at wind-dominant stations. The "Global - No Physics" TFT exhibited the largest drops while the "Global - All - Baseline" model, with its dynamic wind covariates, exhibited the lowest. It is noteworthy that even at the most recalcitrant wind-dominant stations, the degradation of "Global - All - Baseline" was no more than 25% of the station chunk-averaged NRMSE. In other words, late-horizon degradation, while quantifiable, was relatively small. It is noteworthy that in a true forecast mode, in the absence of hindcasted meteorology, the TFT is expected to exhibit more significant late-horizon degradation since meteorological data for the future seven days will incorporate forecast uncertainty.

To summarize these results, in Figure 10 we plot the range of the first-day and sixth-day NRMSE averaged over all available test chunks at each station for both ADCIRC and the "Global - All - Baseline" TFT. Referring to Alaska, it is clear that the TFT was able to render aggressive corrections with minimal long-horizon degradation at the tidally-dominant stations. At the region's four wind-dominant stations, however, the corrections were comparatively weaker and exhibited more significant long-horizon degradation. This same trend was observed in the West and East regions. In the East region, some hybrid stations (tidally-dominant stations with water levels occasionally influenced by winds) exhibited weaker corrections and moderate late-horizon degradation. In the Gulf, while no station exhibited significant late-horizon degradation, no aggressive corrections were rendered save for at the Coast Guard Station in Mobile, Alabama (the TFT dropped the NRMSE from 0.67 to 0.32).

We then considered chunk-averaged performance of the "Global - All - Baseline" TFT at all 228 stations. Here, we grouped chunks by station and then averaged. Refer-

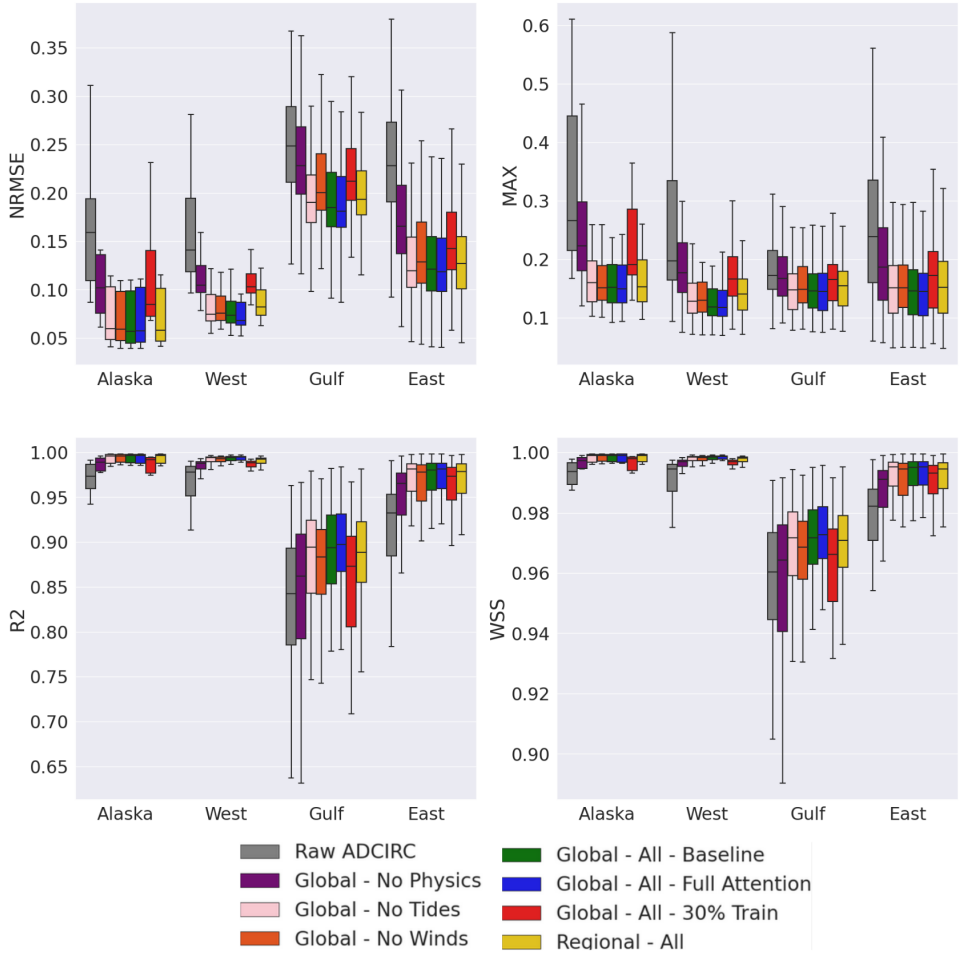

Figure 8. Boxplots of station chunk-averaged evaluation metrics separated by region for ADCIRC and various TFTs. The ML-corrected ADCIRC prediction was considered for each TFT.

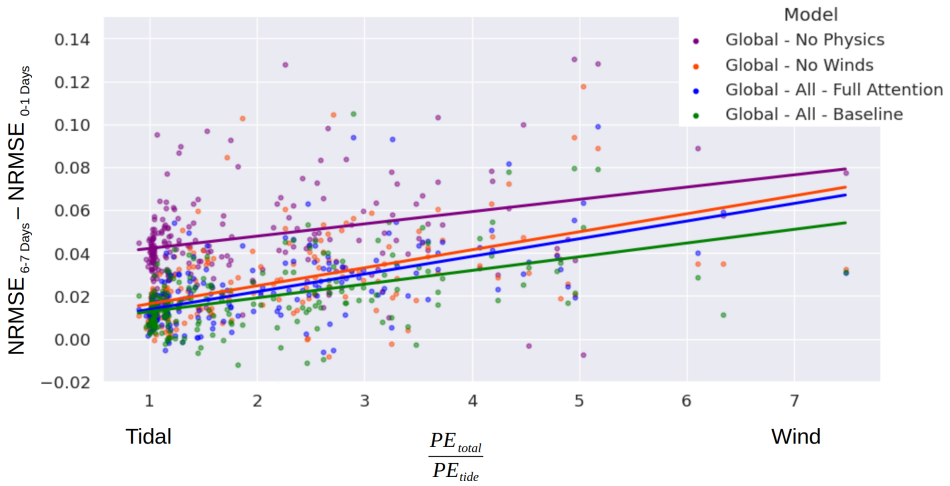

Figure 9. Relationship between late-horizon degradation in ML performance as quantified by $NRMSE_{6-7\ days} - NRMSE_{0-1\ days}$ and tidal-wind dominance as quantified by the ratio between total potential energy and tidal potential energy. In general, every ML model considered indicated late-horizon degradation for both tidally-dominant and wind-dominant stations; however, this degradation was more pronounced at wind-dominant stations. The lines are best-fits to station chunk-averaged degradation for each ML model. All stations were considered.

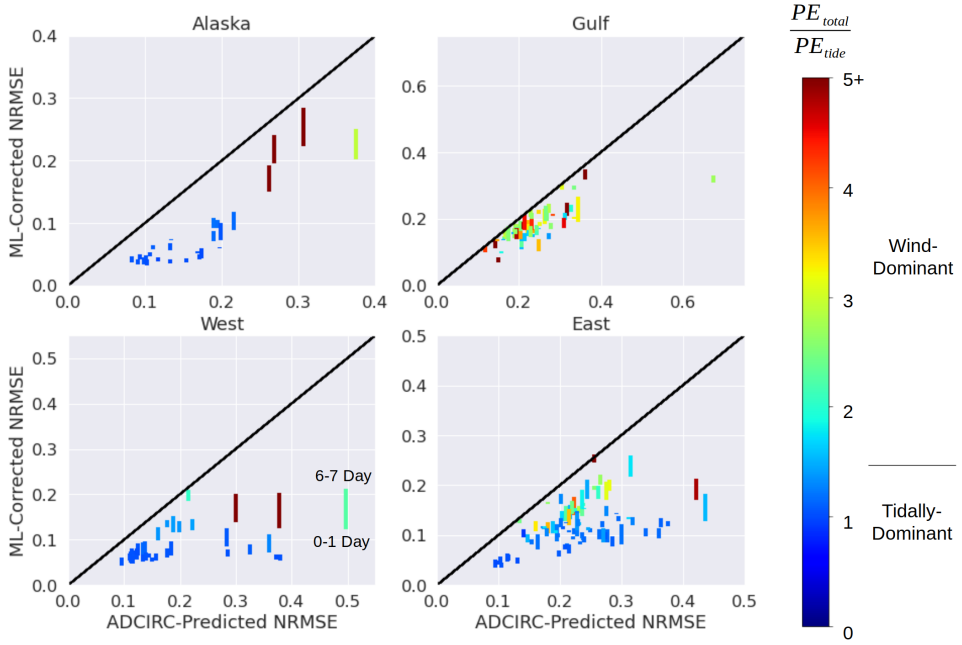

Figure 10. 45-deg plots of the range of ML-corrected vs. ADCIRC-predicted NRMSE for each station. The Global - All - Baseline TFT was used to generate each plot. Each bar corresponds to a station (test chunk-averaged data). The length of each bar indicates the degradation in model skill over the horizon (bottom corresponds to 0-1 day horizon NRMSE, top corresponds to 6-7 day horizon NRMSE). In general, tidally-dominant stations were associated with aggressive corrections and low degradation in skill over the horizon. Wind-dominant stations were associated with comparatively weaker corrections and higher degradation in skill over the horizon.

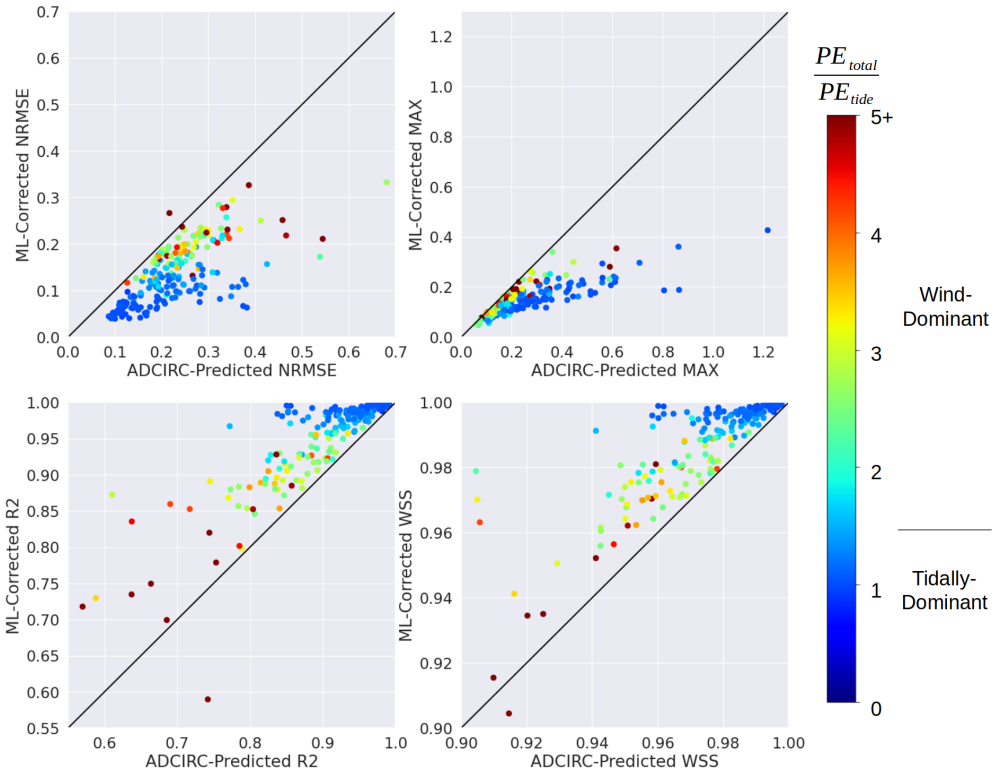

Figure 11. 45-deg plots of ML-corrected vs. ADCIRC-predicted evaluation metrics at the 228 stations throughout the test period. The evaluated ML model was Global - All - Baseline. Each marker corresponds to a station with the metric averaged over all chunks for that station. Colors reference the ratio between a given station's total potential energy to its tidal potential energy, an indication of tidal to wind dominance.

ring to Figure 11, the ML-corrected chunk-averaged station-specific evaluation metrics vs. the ADCIRC-predicted chunk-averaged station-specific evaluation metrics, it is apparent that the TFT was able to render corrections at all but one station (Copano Bay, Texas). The most aggressive corrections occurred at the tidally-dominant stations; 90% of tidal stations saw NRMSE decrease by at least 25% while 50% saw their NRMSE more than halved. The TFT was able to render corrections at wind-dominant stations; however, they were unable to produce the skill observed at the tidally-dominant stations. Regardless, 40% of wind-dominant stations saw their NRMSE decrease by at least 25% while 65% of stations whose ADCIRC NRMSE was greater than 0.4 also saw their NRMSE more than halved.

To clarify this trend further, we investigated six stations in detail: two tidally-dominant (Boston and Anchorage), two wind-dominant (Annapolis, Baltimore), and two riverine (Bridesburg, Pilottown). Figure12 shows 45-deg plots of predicted vs. observed hourly η values for all test chunks for each station for standalone ADCIRC and four TFT-enhanced forecasts with increasing levels of sophistication: the TFT without any physics-based dynamic covariates ("Global - No Physics"), the TFT with all physics-based dynamic covariates except for winds ("Global - No Winds"), the TFT with all covariates ("Global - All - Baseline"), and the region-centric TFTs with all covariates. It is clear that of these four TFTs, "Global - No Physics" was the least performant. It did make modest corrections at the tidally-dominant and riverine stations, suggesting that it picked up on cyclic patterns in the error space via the time covariates, but it was largely unrespon-

sive at wind-dominant stations. It is noteworthy that this model corrected a phasing issue at Bridesburg and partially corrected the anisotropic bias at Pilottown. Broadly, the TFTs appear to be capable of correcting over-damping and under-damping behavior and even phase lags without physics covariates. The "Global - No Winds" TFT was generally more performant with the addition of tidal and ADCIRC covariates. Most notably, it corrected entirely the skewed bias produced by ADCIRC at Pilottown. Adding winds slightly deteriorated performance at the tidally-dominant stations, but significantly improved TFT skill at Annapolis, Baltimore, and Bridesburg. At the tidally-dominant stations, this behavior was likely caused by the TFT attending to winds that were otherwise inconsequential. The inclusion of the tidal covariate was meant to help the TFTs discern the non-tidal contribution of surface water elevations at the stations, but this inclusion did not improve solutions at either tidally-dominant and wind-dominant stations. Including winds in the "Global - All - Baseline" TFT rendered improved performance at Annapolis and Baltimore; however, at the riverine Pilottown station, the addition of winds actually deteriorated performance. Except at the Annapolis station, the region-based models were generally less performant than their global counterpart. This suggests that the TFTs clearly benefited from training on a multitude of signals regardless of region.

To explore the entitlement of adding wind covariates, we plot test-chunk-based NRMSEs for the six stations considered above in Figure 13. Here, we exercised only "Global - All - Baseline" and "Global - No Winds". For the tidally-dominant stations (Boston and Anchorage), it is apparent that both transformers produced predictions of similar skill. Annapolis and Baltimore, both wind-dominant stations, saw marked improvements from adding wind covariates. Moreover, it is clear that large ADCIRC errors were generally associated with high-wind events, suggesting that attending to winds results in enhanced transformer correction capacity. The riverine station Pilottown saw a marginal improvement in chunk-based NRMSE with the inclusion of wind covariates. Finally, the other riverine station, Bridesburg, exhibited an aggressive correction from both TFTs and demonstrated, for a few high-speed wind chunks, the utility of adding wind covariates.

Based on these results, it is apparent that the addition of physics-based dynamic covariates yields enhanced performance at wind-dominant stations. The inclusion of time covariates alone produced considerable improvement at tidally-dominant stations, but failed to render desired performance at the wind-dominant stations. Adding tides did not enhance transformer performance; however, adding ADCIRC's own prediction ("Global - No Winds") further enhanced skill at both station types, but it was only after adding wind covariates that more consistent corrections were rendered at wind-dominant stations (e.g. Annapolis, Baltimore).

4 Conclusions and Discussion

We have coupled the temporal fusion transformer to a high-fidelity global ocean hydrodynamics model, STOFS 2D Global, to render improved station-based predictions of surface water elevation seven days into the future. The STOFS 2D Global model by itself has generally exhibited high skill along US coastlines and adjacent inland coastal waters, but nevertheless, it has systemic model discrepancy stemming from inadequate geometric representation, coarse mesh resolution, incorrect bathymetry, uncertainty in the parameterization of dissipative processes, and meteorological error. Weakly coupling a transformer to this physics-based hydrodynamics model enables the identification of patterns in model error space and their subsequent reduction. We considered several transformers in this study, each supplied with different covariates to ascertain *what* parameters lead to a skillful corrector. We trained and evaluated each transformer on a three-year ADCIRC STOFS 2D Global hindcast. The transformers produced 7-day predictions of ADCIRC error in a "single shot" with 1-hour temporal resolution, ingesting five days

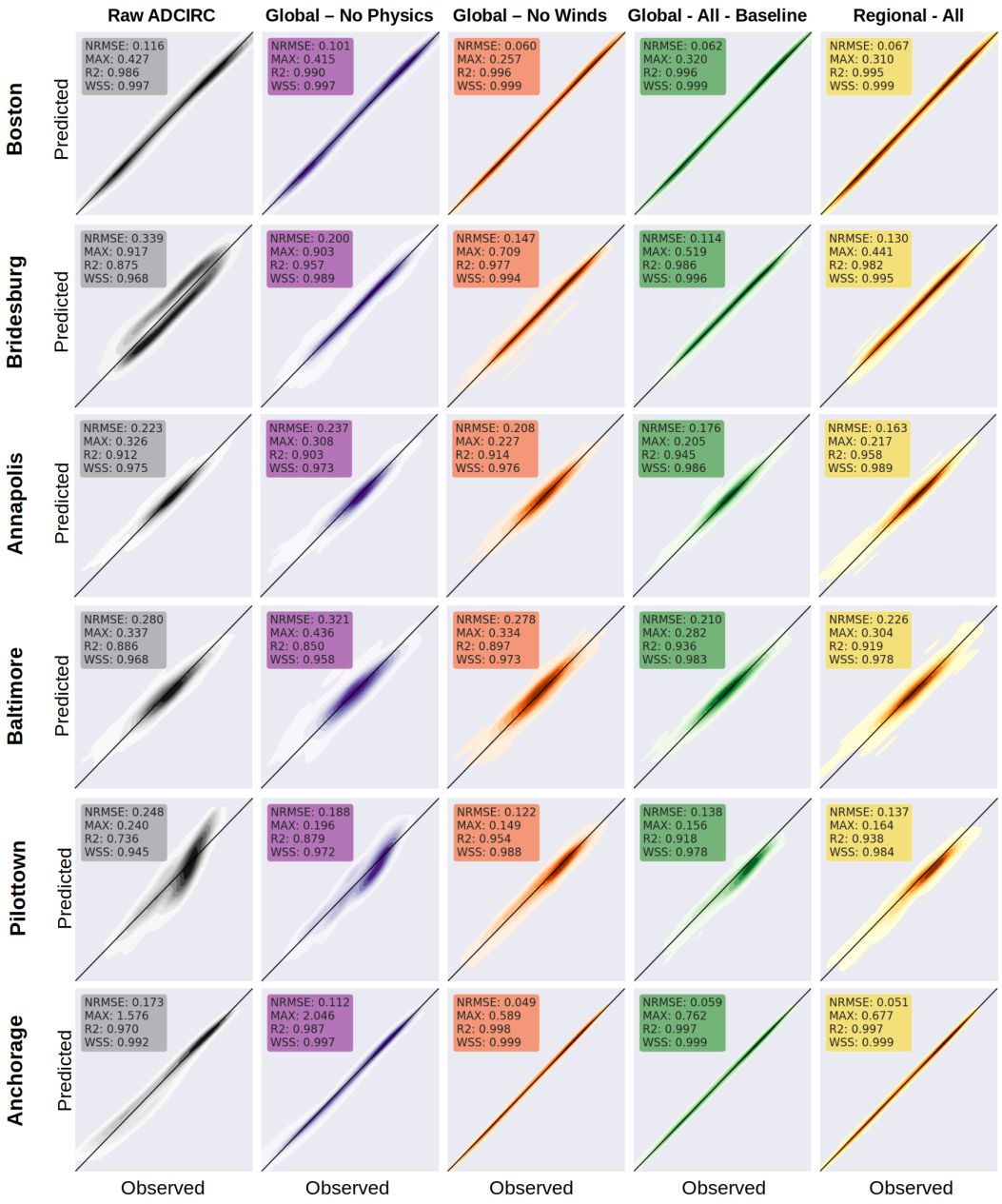

Figure 12. 45-deg plots of predicted vs. observed surface water elevation for hourly data for all test forecast periods. Values from the 6-7 day forecast horizon are used. Annapolis and Baltimore are wind-dominant stations in the Chesapeake Bay (East), Bridesburg is a moderate wind-dominant station located in the Delaware River (East), Anchorage is a tidally dominant station in Alaska, and Pilottown is a wind-dominant station located in Louisiana (Gulf). The ML-corrected ADCIRC prediction was considered for each TFT.

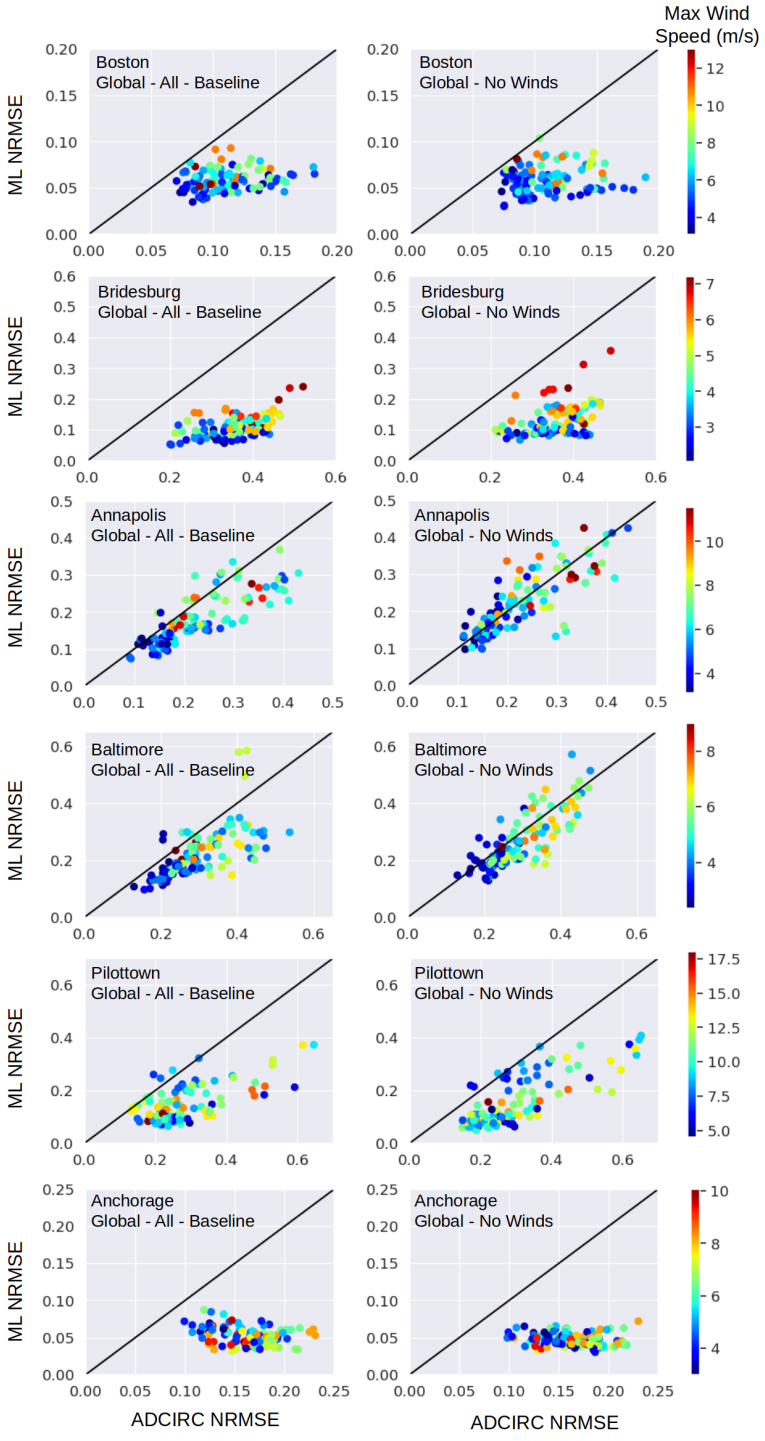

Figure 13. 45-deg plots of ML-corrected vs. ADCIRC-predicted NRMSE averaged for each test chunk at the station. The left column considers Global - All - Baseline, which was trained on winds. The right column considers Global - No Winds, which was not trained on winds. Each marker corresponds to a 7-day output test chunk. In general, the wind-trained transformer was more performant at wind-dominant stations than the transformer trained without winds. Performance at tidally-dominant stations was largely insensitive to winds.

of data prior to the forecast horizon. Based on the evaluation of each transformer, we conclude the following:

- In general, a single transformer exhibits sufficient skill to consistently correct surface water elevations at hundreds of stations along the US coastline.
- Each transformer was particularly capable at tidally-dominant stations. Even in the absence of physics-based dynamic covariates, the TFT was able to render aggressive corrections. For the best-performing transformer, which incorporated ADCIRC computed water levels and wind covariates throughout the 5-day hindcast and 7-day forecast, 50% of tidal stations saw their NRMSE (averaged over the test period) halved. In certain cases, the inclusion of wind covariates slightly deteriorated performance at tidally-dominant stations. This suggests that the TFT was attending to winds, even in cases when ADCIRC error was wind-invariant.
- At wind-dominant stations, adding more physics-based dynamic covariates led to enhanced skill. Including none of these resulted in little to no corrections. Adding ADCIRC's predictions certainly improved performance at wind-dominant stations; however, wind covariates were necessary to correct recalcitrant test chunks, especially those with high wind-speed events. With the full complement of physics-based covariates, 40% of wind-dominant stations saw their NRMSE reduced by at least 25%.
- Transformers were either trained on all stations along US coastlines or on stations in specific regions. Region-centric transformers exhibited slightly diminished performance compared to their counterparts trained on all available stations. The wind-dominant Annapolis station was one exceptional station wherein the region-centric model out-performed the global transformers.
- Of all the transformers considered, a TFT trained on 2-years of the hindcast with the full complement of past, future, and static covariates was, with a few exceptions, the most performant. This TFT was denoted "Global - All - Baseline" herein. Exceptions include the aforementioned slight deterioration at tidally-dominant stations (owing to needlessly attending to winds) and rare instances of region-based TFT superiority. The TFT trained with full attention, so that its decoder could attend to forthcoming future covariates, exhibited roughly similar performance with a larger training cost as compared to the same model whose decoder could only attend to previous future covariates. Both of these models had encoders that did attend to both past and future covariates.
- The TFTs exhibited modest amounts of late-horizon degradation in skill. This degradation was measured, test chunk by test chunk, as the difference between the NRMSE for the 6-7 day horizon and the NRMSE for the 0-1 day horizon. This degradation was noted to be the highest at some wind-dominant stations; however, it never exceeded 25% of a given station's average NRMSE.

The approach proposed herein is station-centric. Training and evaluation data was mapped to hundreds of stations situated along US coastlines, and improvements were rendered by the transformers at these specific locations. The extrapolative capability of the TFT was not assessed. While it could theoretically be used to extrapolate beyond the trained stations, it is likely that the training set size would need to include thousands (and not hundreds) of stations so that the TFT could draw correlations between the supplied location-based static covariates, the physics-based dynamic covariates, and the target signal.

The framework was challenged to attend to high wind-speed events. In the present work, the vast majority of chunks did not incorporate elevated wind levels and their associated elevated surges. The few chunks that did were not accommodated in any particular fashion. In fact, while MSE loss penalized outliers aggressively, it did so in batches. Consequently, the few chunks with fringe events were likely muted by the more frequent

quiescent chunks. There are potential pathways to circumvent this undesirable behavior. For example, in imbalanced classification problems, class weights make the optimizer more cognizant of under-represented classes. By tagging our fringe chunks and amplifying their loss, we hypothesize that the optimizer may become more sensitive to them. However, this does not address the shortcomings of the wind product. CFSv2 has 0.25 deg spatial resolution. Consequently, it is not expected to perform well in regions with variable terrain and complex shallow inland water systems whose water levels are particularly susceptible to strong winds. Additionally, this lack of resolution means that it cannot adequately resolve strong frontal systems and high-energy, low-pressure wind events such as tropical cyclones which tend to be muted in the product. Finally, it does not have skillful inland atmospheric boundary layer adjustments, further reducing its utility for the inland and near-shore stations considered herein.

Finally, maturing the proposed framework for operational forecasting will require adopting forecasted meteorology and water level predictions in place of hindcasted values. This will introduce a greater level of epistemic uncertainty. Herein, we trained our TFTs using a deterministic loss function. This deterministic model could be used to facilitate forward propagation in an ensemble uncertainty quantification scheme. For example, each member of NOAA's Global Ensemble Forecast System (GEFS) could be forced independently through the transformer. Thereafter, probabilities of exceedance could be calculated at each station. Additionally, leveraging quantile regression in place of a deterministic loss function could help quantify confidence of each correction in time. Collectively, this strategy offers a compelling pathway to operations which are becomingly increasingly stochastic in nature.

Open Research Section

Results from the three-year ADCIRC STOFS 2D Global hindcast used to train the transformers considered herein can be downloaded from the "Improving Storm Surge Forecasts with Transformers" project (Cerrone et al., 2023) on DesignSafe (NSF-NHERI, 2023). Moreover, output from each transformer can be found under the same DesignSafe project. This project can be accessed here: `https://doi.org/10.17603/ds2-t5mf-3757`

Acknowledgments

The authors would like to acknowledge Drs. Rick Luettich and Shintaro Bunya, both of the University of North Carolina at Chapel Hill, and Dr. Eirik Valseth, of the Norwegian University of Life Sciences and the Oden Institute for Computational Engineering and Sciences at the University of Texas at Austin, for their valuable feedback during the early stages of this work. This work was funded by the United States Department of Energy (DOE) award DE-SC0022316. The three-year hindcast was run on TACC's Frontera supercomputer with allocation "High Fidelity Hurricane Storm Surge and Ocean Modeling" (DMS21031).

References

Adcirc. (2023). Retrieved 22.11.2023, from `\url{https://adcirc.org/}`

Ayyad, M., Hajj, M. R., & Marsooli, R. (2022, November). Machine learning-based assessment of storm surge in the new york metropolitan area. *Sci. Rep.*, *12*(1), 1–12.

Beaman, R. (2018). High-resolution depth model for the northern australia—100 m. *Geoscience Australia, Canberra*.

Blakely, C. P., Ling, G., Pringle, W. J., Contreras, M. T., Wirasaet, D., Westerink, J. J., ... Massey, C. (2022, May). Dissipation and bathymetric sensitivities in an unstructured mesh global tidal model. *J. Geophys. Res. C: Oceans*,

127(5).

Bolton, T., & Zanna, L. (2019, January). Applications of deep learning to ocean data inference and subgrid parameterization. *J. Adv. Model. Earth Syst.*, *11*(1), 376–399.

Bonavita, M., & Laloyaux, P. (2020, December). Machine learning for model error inference and correction. *Journal of Advances in Modeling Earth Systems*, *12*(12), e2020MS002232.

Bunya, S., Dietrich, J. C., Westerink, J. J., Ebersole, B. A., Smith, J. M., Atkinson, J. H., ... Roberts, H. J. (2010, February). A High-Resolution coupled riverine flow, tide, wind, wind wave, and storm surge model for southern louisiana and mississippi. part i: Model development and validation. *Mon. Weather Rev.*, *138*(2), 345–377.

Butler, T., Altaf, M. U., Dawson, C., Hoteit, I., Luo, X., & Mayo, T. (2012). Data assimilation within the advanced circulation (adcirc) modeling framework for hurricane storm surge forecasting. *Monthly Weather Review*, *140*(7), 2215 - 2231. Retrieved from `https://journals.ametsoc.org/view/journals/mwre/140/7/mwr-d-11-00118.1.xml` doi: https://doi.org/10.1175/MWR-D-11-00118.1

Butler, T., Graham, L., Estep, D., Dawson, C., & Westerink, J. (2015). Definition and solution of a stochastic inverse problem for the manning's n parameter field in hydrodynamic models. *Advances in Water Resources*, *78*, 60-79. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0309170815000135` doi: https://doi.org/10.1016/j.advwatres.2015.01.011

Cerrone, A., Westerink, L., Dawson, C. N., & Westerink, J. (2023). *Improving storm surge forecasts with transformers.* [Dataset] Designsafe-CI. Retrieved from `https://www.designsafe-ci.org/data/browser/public/designsafe.storage.published/PRJ-4200` doi: 10.17603/DS2-T5MF-3757

Chen, C., Beardsley, R. C., Luettich Jr., R. A., Westerink, J. J., Wang, H., Perrie, W., ... Toulany, B. (2013). Extratropical storm inundation testbed: Intermodel comparisons in scituate, massachusetts. *Journal of Geophysical Research: Oceans*, *118*(10), 5054-5073. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/jgrc.20397` doi: https://doi.org/10.1002/jgrc.20397

De Kleermaeker, S., Verlaan, M., Mortlock, T., Rego, J., Irazoqui, M., Yan, K., & Twigt, D. (2017, 06). Global-to-local scale storm surge modelling on tropical cyclone affected coasts..

de Oliveira, M. M. F., Ebecken, N. F. F., de Oliveira, J. L. F., & de Azevedo Santos, I. (2009, January). Neural network model to predict a storm surge. *J. Appl. Meteorol. Climatol.*, *48*(1), 143–155.

Dietrich, J. C., Westerink, J. J., Kennedy, A. B., Smith, J. M., Jensen, R. E., Zijlema, M., ... Cobell, Z. (2011). Hurricane gustav (2008) waves and storm surge: Hindcast, synoptic analysis, and validation in southern louisiana. *Monthly Weather Review*, *139*(8), 2488 - 2522. Retrieved from `https://journals.ametsoc.org/view/journals/mwre/139/8/2011mwr3611.1.xml` doi: https://doi.org/10.1175/2011MWR3611.1

Elfring, J., Torta, E., & van de Molengraft, R. (2021, January). Particle filters: A Hands-On tutorial. *Sensors*, *21*(2).

Fisheries, & Canada, O. (2023). *Canadian Hydrographic Service Non-Navigational (NONNA) Bathymetric Data [10m resolution].* Abacus Data Network. Retrieved from `https://hdl.handle.net/11272.1/AB2/YJPER2` doi: 11272.1/AB2/YJPER2

Hager, W. H., Castro-Orgaz, O., & Hutter, K. (2019). Correspondence between de saint-venant and boussinesq. 1: Birth of the shallow–water equations. *Comptes Rendus Mécanique*, *347*(9), 632-662. Retrieved from `https://www.sciencedirect.com/science/article/pii/S1631072119301263` doi:

https://doi.org/10.1016/j.crme.2019.08.004

Haidvogel, D. B., Arango, H. G., Hedstrom, K., Beckmann, A., Malanotte-Rizzoli, P., & Shchepetkin, A. F. (2000). Model evaluation experiments in the north atlantic basin: simulations in nonlinear terrain-following coordinates. *Dynamics of Atmospheres and Oceans*, *32*(3), 239-281. Retrieved from `https://www.sciencedirect.com/science/article/pii/S037702650000049X` doi: https://doi.org/10.1016/S0377-0265(00)00049-X

Hervouet, J.-M. (2007). *Hydrodynamics of free surface flows: Modelling with the finite element method*. Chichester, West Sussex: John Wiley & Sons Ltd.

Herzen, J., Lässig, F., Piazzetta, S. G., Neuer, T., Tafti, L., Raille, G., ... Grosch, G. (2022). Darts: User-Friendly modern machine learning for time series. *J. Mach. Learn. Res.*, *23*(124), 1–6.

Holland, M. (2020). *An introduction to the extended kalman filter*. Nova Science Publishers.

Hope, M. E., Westerink, J. J., Kennedy, A. B., Kerr, P. C., Dietrich, J. C., Dawson, C., ... Westerink, L. G. (2013, September). Hindcast and validation of hurricane ike (2008) waves, forerunner, and storm surge. *J. Geophys. Res. C: Oceans*, *118*(9), 4424–4460.

IHO-UNESCO. (2020). *Gebco_2020 grid* (No. DOI: 10.5285/a29c5465-b138-234d-e053-6c86abc040b9). (`https://www.gebco.net/data_and_products/gridded_bathymetry_data/gebco_2020/`)

Joyce, B. R., Gonzalez-Lopez, J., Van der Westhuysen, A. J., Yang, D., Pringle, W. J., Westerink, J. J., & Cox, A. T. (2019). U.s. ioos coastal and ocean modeling testbed: Hurricane-induced winds, waves, and surge for deep ocean, reef-fringed islands in the caribbean. *Journal of Geophysical Research: Oceans*, *124*(4), 2876-2907. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018JC014687` doi: https://doi.org/10.1029/2018JC014687

Joyce, B. R., Pringle, W. J., Wirasaet, D., Westerink, J. J., Van der Westhuysen, A. J., Grumbine, R., & Feyen, J. (2019). High resolution modeling of western alaskan tides and storm surge under varying sea ice conditions. *Ocean Modelling*, *141*, 101421. Retrieved from `https://www.sciencedirect.com/science/article/pii/S1463500318303834` doi: https://doi.org/10.1016/j.ocemod.2019.101421

Kerr, P. C., Donahue, A. S., Westerink, J. J., Luettich Jr., R. A., Zheng, L. Y., Weisberg, R. H., ... Cox, A. T. (2013). U.s. ioos coastal and ocean modeling testbed: Inter-model evaluation of tides, waves, and hurricane surge in the gulf of mexico. *Journal of Geophysical Research: Oceans*, *118*(10), 5129-5172. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/jgrc.20376` doi: https://doi.org/10.1002/jgrc.20376

Kerr, P. C., Martyr, R. C., Donahue, A. S., Hope, M. E., Westerink, J. J., Luettich Jr., R. A., ... Westerink, H. J. (2013). U.s. ioos coastal and ocean modeling testbed: Evaluation of tide, wave, and hurricane surge response sensitivities to mesh resolution and friction in the gulf of mexico. *Journal of Geophysical Research: Oceans*, *118*(9), 4633-4661. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/jgrc.20305` doi: https://doi.org/10.1002/jgrc.20305

Laplace, P. (1776). Recherches sur plusieurs points du système du monde. *Mémoires de l'Académie (royale) des sciences de l'Institut (imperial) de France*, *89*, 177–264.

Li, C., Mahadevan, S., Ling, Y., Wang, L., & Choze, S. (2017, January). A dynamic bayesian network approach for digital twin. In *19th AIAA Non-Deterministic approaches conference.* Reston, Virginia: American Institute of Aeronautics and Astronautics.

Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2021). *Temporal fusion transformers for interpretable multi-horizon time series forecasting* (Vol. 37) (No. 4).

Lyons, M., Larsen, K., & Skone, M. (2022, June). *CoralMapping/AllenCoralAtlas: DOI for paper at v1.3.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.6622015 doi: 10.5281/zenodo.6622015

NOAA-EMC. (2023). *The global forecast system.* Retrieved from https://www.emc.ncep.noaa.gov/emc/pages/numerical_forecast_systems/gfs.php

NOAA-NOS. (2023). *Storm surge & tide operational forecast system.* Retrieved from https://polar.ncep.noaa.gov/estofs/global/index.htm

NOAA-OPC. (2023). *Global stofs atlantic region storm surge model guidance.* Retrieved from https://ocean.weather.gov/estofs/estofs_surge_twlev.php

NSF-NHERI. (2023). *Designsafe.* Retrieved 22.11.2023, from https://www.designsafe-ci.org/

Pachev, B., Arora, P., del Castillo-Negrete, C., Valseth, E., & Dawson, C. (2023). A framework for flexible peak storm surge prediction. *Coastal Engineering*, *186*, 104406. Retrieved from https://www.sciencedirect.com/science/article/pii/S0378383923001308 doi: https://doi.org/10.1016/j.coastaleng.2023.104406

Pe'eri, S. (2023). *Updated: Upgrade of the surge and tide operational forecast system (stofs, formerly estofs) to version 1.1.0: Effective january 10, 2023* (Tech. Rep.). NOS/Office of Coast Survey. Retrieved from https://www.weather.gov/media/notification/pdf2/scn22-108_stofs_v1.1.0_aab.pdf

Pringle, W. J., Wirasaet, D., Roberts, K. J., & Westerink, J. J. (2021, February). Global storm tide modeling with ADCIRC v55: unstructured mesh design and performance. *Geoscientific Model Development*, *14*(2), 1125–1145.

Pringle, W. J., Wirasaet, D., Suhardjo, A., Meixner, J., Westerink, J. J., Kennedy, A. B., & Nong, S. (2018). Finite-element barotropic model for the indian and western pacific oceans: Tidal model-data comparisons and sensitivities. *Ocean Modelling*, *129*, 13-38. Retrieved from https://www.sciencedirect.com/science/article/pii/S146350031830026X doi: https://doi.org/10.1016/j.ocemod.2018.07.003

Ristic, B., Arulampalam, S., & Gordon, N. (2003). *Beyond the kalman filter: Particle filters for tracking applications.* Artech House.

Roelvink, J. A., & van Banning, G. (1995). Design and development of delft3d and application to coastal morphodynamics. *Oceanographic Literature Review*, *11*, 925. Retrieved from https://api.semanticscholar.org/CorpusID:127930966

Rougier, J., Brady, A., Bamber, J., Chuter, S., Royston, S., Vishwakarma, B. D., ... Ziegler, Y. (2023). The scope of the kalman filter for spatio-temporal applications in environmental science. *Environmetrics*, *34*(1), e2773. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/env.2773 doi: https://doi.org/10.1002/env.2773

Schaffer, J., Timmermann, R., Arndt, J. E., Kristensen, S. S., Mayer, C., Morlighem, M., & Steinhage, D. (2016). A global, high-resolution data set of ice sheet topography, cavity geometry, and ocean bathymetry. *Earth System Science Data*, *8*(2), 543–557. Retrieved from https://essd.copernicus.org/articles/8/543/2016/ doi: 10.5194/essd-8-543-2016

Seroka, G., Funakoshi, Y., Yang, Z., Moghimi, S., Myers, E., Pe'eri, S., ... Kaiser, C. (2023, January). Upgrades to NOAA/NOS' surge and tide operational forecast system (STOFS). In *103rd AMS annual meeting.* AMS.

Stammer, D., Ray, R. D., Andersen, O. B., Arbic, B. K., Bosch, W., Carrère, L., ... Yi, Y. (2014). Accuracy assessment of global barotropic ocean tide models. *Reviews of Geophysics*, *52*(3), 243-282. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014RG000450 doi:

https://doi.org/10.1002/2014RG000450

Tiggeloven, T., Couasnon, A., van Straaten, C., Muis, S., & Ward, P. J. (2021, August). Exploring deep learning capabilities for surge predictions in coastal areas. *Sci. Rep.*, *11*(1), 17224.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *CoRR*, *abs/1706.03762*. Retrieved from http://arxiv.org/abs/1706.03762

Verlaan, M., De Kleermaeker, S., & Buckman, L. (2015). Glossis: Global storm surge forecasting and information system. In *Australasian coasts & ports conference 2015: 22nd australasian coastal and ocean engineering conference and the 15th australasian port and harbour conference.* Auckland, New Zealand. Retrieved from https://search.informit.org/doi/10.3316/informit.703696922952912

Wang, P., & Bernier, N. B. (2023). Adding sea ice effects to a global operational model (nemo v3.6) for forecasting total water level: approach and impact. *Geoscientific Model Development*, *16*(11), 3335–3354. Retrieved from https://gmd.copernicus.org/articles/16/3335/2023/ doi: 10.5194/gmd-16-3335-2023

Wang, P., Bernier, N. B., Thompson, K. R., & Kodaira, T. (2021). Evaluation of a global total water level model in the presence of radiational s2 tide. *Ocean Modelling*, *168*, 101893. Retrieved from https://www.sciencedirect.com/science/article/pii/S1463500321001463 doi: https://doi.org/10.1016/j.ocemod.2021.101893

Westerink, J. J., Luettich, R. A., Feyen, J. C., Atkinson, J. H., Dawson, C., Roberts, H. J., ... Pourtaheri, H. (2008, March). A basin- to Channel-Scale unstructured grid hurricane storm surge model applied to southern louisiana. *Mon. Weather Rev.*, *136*(3), 833–864.

Willmott, C. J. (1981). On the validation of models. *Physical Geography*, *2*(2), 184-194. Retrieved from https://doi.org/10.1080/02723646.1981.10642213 doi: 10.1080/02723646.1981.10642213

Wood, D. (2023). *Global storm and tide operational forecast system - global stofs.* Notre Dame Computational Hydraulics Laboratory. Retrieved from https://dylnwood.github.io/GESTOFS-develop/

Xie, W., Xu, G., Zhang, H., & Dong, C. (2023). Developing a deep learning-based storm surge forecasting model. *Ocean Modelling*, *182*, 102179. Retrieved from https://www.sciencedirect.com/science/article/pii/S1463500323000203 doi: https://doi.org/10.1016/j.ocemod.2023.102179

Zampieri, L., Arduini, G., Holland, M., Keeley, S. P. E., Mogensen, K., Shupe, M. D., & Tietsche, S. (2023, May). A machine learning correction model of the winter Clear-Sky temperature bias over the arctic sea ice in atmospheric reanalyses. *Mon. Weather Rev.*, *151*(6), 1443–1458.