TidalMark: a Scalable Benchmark for Coastal Water Level Forecasting

Lucas A. Raicu
University of Chicago
Chicago, IL, USA
lucas.raicu@gmail.com

Daniel Grzenda, Ian Foster, Kyle Chard (Advisors)
University of Chicago
Chicago, IL, USA
{grzenda,foster,chard}@uchicago.edu

ABSTRACT

Accurate forecasting of water levels is essential for flood mitigation. Traditionally, predictions have been based on harmonic analysis and sensor networks maintained by the National Oceanographic and Atmospheric Administration. However, these methods struggle with high-variance events that change water levels from the long-term tidal baseline. TidalMark evaluates the ability of a variety of deep learning models to model these high-variance events. Through extensive hyperparameter sweeps and comparisons across model variants, we have evaluated tradeoffs in accuracy, generalization, and scalability. Our results show that properly tuned machine learning models consistently outperform the scientific-standard harmonic approaches between 2.1X and 4.7X (between one to seven day predictions) with the goal towards achieving adaptive, scalable, and accurate forecasting of coastal water levels.

KEYWORDS

Water level forecasting, LSTM, deep learning, flood prediction, time-series modeling

ACM Reference Format:

Lucas A. Raicu and Daniel Grzenda, Ian Foster, Kyle Chard (Advisors). 2025. TidalMark: a Scalable Benchmark for Coastal Water Level Forecasting. In . ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Coastal communities face increasing threats from flooding, sea-level rise, and extreme weather [5]. Since the 1800s, harmonic analysis has been used to predict water levels by decomposing tides into cyclical components [1, 3]. The National Oceanographic and Atmospheric Administration (NOAA) [2] maintains hundreds of sensors that measure coastal water levels and use harmonic analysis to predict future water levels. While effective under stable conditions, harmonic analysis assumes linearity and stationarity, limiting accuracy during environmental variability. Figure 1 shows the water level during a major weather system over the course of several days that produced flooding water levels (red line). Unfortunately, both NOAA predictions and Forecast Guidance severely underestimated the water level.

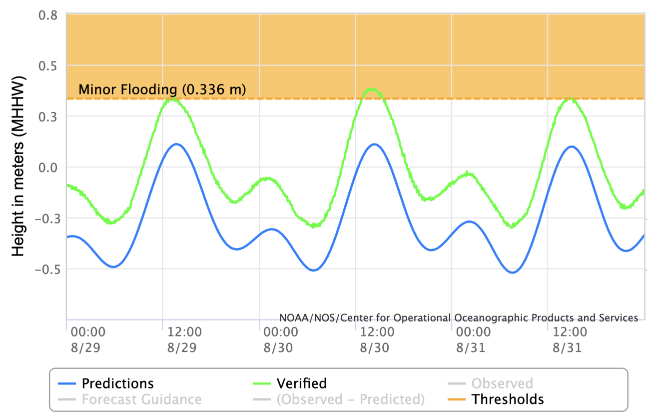


Figure 1: NOAA's largest prediction error for the Nawiliwili station in 2024

NOAA determines harmonic constituents by collecting and analyzing more than 21 years of continuous station data, extracting the tidal frequencies, amplitudes, and phases through long-term statistical averaging. In reality, coastal systems exhibit non-stationary behavior (e.g., long-term sea-level change) and nonlinear weather interactions (e.g., storm surge), which harmonic analysis cannot adapt to dynamically.

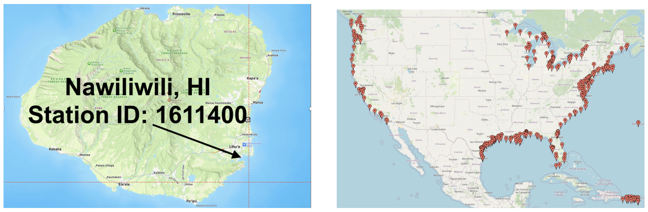
We propose to develop TidalMark, a benchmark that applies deep learning models to improve forecast accuracy in coastal water-level prediction. Our main contribution is comparison of cutting-edge water level prediction models on a novel dataset.

2 WATER LEVEL

Our dataset originates from NOAA's National Water Level Observation Network (NWLON), spanning 217 stations. Over a five-year period from 2019-2024 the dataset has over 127 million measurements, each taken at six-minute intervals. Our results focus on the station in Nawiliwili, HI (Station ID: 1611400) due to the data being complete with no missing samples, see Figure 2a).

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn



(a) NOAA's Nawiliwili station (b) U.S. Map of Stations

Figure 2: Target station and dataset coverage

We developed a modular pipeline in PyTorch for training and evaluating water-level forecasting models. Each model receives a fixed-length window of prior water levels (7 or 14 days) as input and predicts future water levels at multiple time horizons (1, 3, 5, and 7 days). To understand the importance of exogenous variables in accurate water level forecasting, we examined both univariate and multivariate models. We tested standard Long-Short Term Memory (LSTM) models, bidirectional LSTMs (BiLSTM), convolutional LSTMs (Conv-LSTM), attention-based LSTMs (Attn-LSTM), as well as single and stacked gated recurrent units (GRUs). See Table 1 for hyperparameter grid search values.

Table 1: Hyperparameter grid for univariate model sweep.

Parameter	Values Tested
Sequence Length	7, 14
Batch Size	32, 64, 128, 256, 512
Learning Rate	$1\text{x}10^{-3}$, $1\text{x}10^{-4}$, $1\text{x}10^{-5}$
Hidden Size	32, 64
Number of Layers	1, 2

3 PERFORMANCE EVALUATION

Despite the theoretical advantages of BiLSTM, GRUs, and Attention-based LSTMs, we did not observe a significant improvement over standard LSTM models (see Table 2). Performance distributions largely overlapped, suggesting that careful tuning matters more than exotic architecture selection.

Table 2: Forecast performance across architectures

Model	Seq. Len	Batch Size	LR	Hidden Size	Layers	MSE	MAE	MaxAE
ATTN-LSTM	7	32	0.001	64	1	0.0033	0.0441	0.3085
BiLSTM	7	32	0.001	32	1	0.0029	0.0409	0.2758
Conv-LSTM	7	64	0.001	32	1	0.0028	0.0400	0.2556
GRU	7	64	0.0001	64	2	0.0028	0.0402	0.2709
LSTM	7	128	0.0005	32	1	0.0028	0.0400	0.3423

For the hyperparamter sweep, we found learning rate to be the dominant factor. Models trained with 1×10^{-3} converged fastest and achieved the highest R^2, outperforming lower values by a wide margin. Batch size and number of layers had modest effects. Smaller batches (32–64) improved stability and generalization. One or two layers offered comparable results.

Longer sequences (14 vs. 7) slightly improved accuracy but at a significant training time increase. Models tuned for one sequence length also did not generalize well for other sequence lengths.

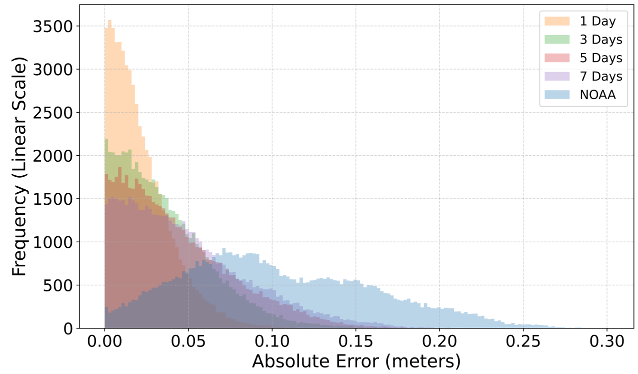


Figure 3: LSTM vs NOAA prediction errors (Nawiliwili station, 2024). LSTM errors provided at 1, 3, 5, and 7 day horizons.

Our best configuration for the LSTM had 1-layer, input sequence of 14-days, 64 batch size, 32 hidden size, and $1\text{x}10^{-3}$ learning rate. Our results show that properly tuned machine learning models outperform the scientific standard harmonics (2.1X with 7 day horizon or 4.7X with 1 day horizon). Figure 3 shows the histogram of all predictions using the LSTM compared to NOAA's predictions, showing the error in meters. These findings align with broader trends in Earth system modeling, where deep learning methods are increasingly surpassing classical statistical and physical models in accuracy and adaptability [4].

4 CONCLUSION AND FUTURE WORK

TidalMark demonstrates that well-tuned deep learning architectures can outperform traditional harmonic-based forecasts in dynamic environments. Our results show that properly tuned machine learning models outperform state-of-the-art approaches by 4.7X. We plan to extend TidalMark into a full spatiotemporal Graph Neural Network framework, where each station is a node and edges represent geophysical relationships. Beyond methodological advances, we plan to integrate TidalMark into real-time early flood detection and coastal management systems.

REFERENCES

[1] Sergio Consoli, Diego Reforgiato Recupero, and Vanni Zavarella. 2014. A survey on tidal analysis and forecasting methods for Tsunami detection. *arXiv preprint arXiv:1403.0135* (2014). Traditional tidal forecasting methods based on harmonic analysis require long-term data and fail under non-astronomical influences such as weather.

[2] National Oceanic and Atmospheric Administration (NOAA). 2025. National Oceanic and Atmospheric Administration. https://www.noaa.gov/. Accessed: 2025-08-08.

[3] David Pugh. 2004. *Tides, Surges and Mean Sea-Level* (2nd ed.). Wiley.

[4] Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and Prabhat. 2019. Deep learning and process understanding for data-driven Earth system science. *Nature* 566, 7743 (2019), 195–204.

[5] NOAA Ocean Service. 2024. What threats do coastal communities face? Online facts page. Lists threats including extreme natural events, sea level rise and coastal erosion..