A Transformer-Based Regression Scheme for Forecasting Significant Wave Heights in Oceans

Pujan Pokhrel, Elias Ioup, Julian Simeonov, Md Tamjidul Hoque, and Mahdi Abdelguerfi

***Abstract*—In this article, we present a novel approach for forecasting significant wave heights in oceanic waters. We propose an algorithm based on the WaveWatch III, differencing, and a transformer neural network (Transformer). The data becomes stationary after first-order differencing, performed with the observed significant wave height and the wave height forecasts obtained from WaveWatch III. We perform a case study on a group of 92 buoys using WaveWatch III hindcasts. The Transformer model then provides the statistical forecasts of the residuals. The Transformer-based proposed framework obtains the root mean square error of 0.231 m for two days ahead forecasting. Our proposed method outperforms existing state-of-the-art machine learning and numerical approaches for significant wave heights prediction. Our results suggest that combining numerical and machine learning approaches gives better performance than using either alone.**

***Index Terms*—Data assimilation, forecasting, residual correction, significant wave heights, Transformer.**

I. Introduction

THERE has been a growing interest in understanding the physics behind waves and their dynamics in weakly nonlinear media like hydrodynamics [1], optics [2], quantum mechanics [3], Bose–Einstein condensates [4], [5], and finance [6]. These media sometimes experience large waves, often known as rogue waves, and are poorly understood in terms of their formation and dynamics. So far, the main way of studying these waves is through classical nonlinear equations [7] which include the nonlinear Schrodinger equation [8], [9], Korteweg–de Vries (kDV) equation [10], Kadomtsev–Petviashvili equation [11], Zakharov equation [12], and fully nonlinear potential systems [13]. Extending these equations to macroscopic systems is difficult because it is computationally time-intensive. In recent years, since point measurements over a long duration of time are available for ocean waves through buoys, several efforts have been made to understand wave dynamics in oceans. These waves show analogy in other wave media [14] and, thus, help elucidate the factors behind wave formation. Pokhrel *et al.* [15] have used statistical machine learning (ML) methods to forecast anomalous waves through binary predictions on whether the wave is large or not. However, to study these anomalous waves, it is important to understand the dynamics of the height distribution of the waves via regression approaches.

This article focuses on the accurate prediction of ocean waves' significant wave heights, which remains one of the most critical outstanding classical physics problems [16]. While ocean waves are used to study the analogies in various media, nowcasting and forecasting of waves are also crucial for myriad other reasons, including optimizing ship routes for efficient shipping, avoiding disasters, aiding the aquaculture industry, safely conducting military and amphibious operations by the Navy and Marine Corps teams, etc. The other importance of wave prediction lies in the efficient renewable energy generated from renewable energy sources like solar, wind, tidal, wave, etc.

There are previous ML and statistical methods applied to forecast significant wave heights [17]–[23]. While these methods have provided substantial insights into the problem, accurate prediction remains elusive. There also exist deterministic approaches for forecasting significant wave heights. These methods take nonlinear equations like nonlinear Schrodinger equation [24], [25], kDV equation [26], or Zakharov equation [27] and then use a spectral method based on fast Fourier transform (FFT) to perform deterministic forecasts on unidirectional wave fields. The numerical methods must be reinitialized for various geographical areas [28] with different conditions for forecasting. However, ML methods are easily generalizable to various domains [29] but are not competitive with the numerical methods for forecasting significant wave heights. The performances of the ML methods are reported on individual buoys [18]–[20], [29], [30]. Since existing ML methods [18], [20], [30] are trained on single buoys, their generalized performance on a group of buoys is difficult to calculate.

Various studies have previously used deep learning methods to forecast significant wave heights [31], [32]. Wei [33] used an artificial neural network (ANN) to build an AI-based storm forecasting system to show that with the sliding window size equal to the forecast horizon, decent prediction accuracy can be achieved. Pirhoshyaran and Snyder [31] used recurrent and

Manuscript received April 6, 2021; revised November 29, 2021; accepted April 29, 2022. This work was supported by the U.S. Navy (Office of Naval Research) under Contract N00173-16-2-C902. *(Corresponding author: Md Tamjidul Hoque.)*

Associate Editor: M. Haller.

Pujan Pokhrel, Md Tamjidul Hoque, and Mahdi Abdelguerfi are with the Canizaro Livingston Gulf States Center for Environmental Informatics, University of New Orleans, New Orleans, LA 70148 USA (e-mail: ppokhrel@uno.edu; thoque@uno.edu; chairman@cs.uno.edu).

Elias Ioup is with the Center for Geospatial Sciences Naval Research Laboratory Stennis Space Center Mississippi, Mississippi, MS 39529 USA (e-mail: elias.ioup@nrlssc.navy.mil).

Julian Simeonov is with the Ocean Sciences Division Naval Research Laboratory Stennis Space Center Mississippi, Mississippi, MS 39529 USA (e-mail: julian.simeonov@nrlssc.navy.mil).

Digital Object Identifier 10.1109/JOE.2022.3173454

sequence-to-sequence networks to show that recurrent structures with deep networks have better prediction accuracy than shallow networks. Moreover, the authors also introduced the concept of "refined buoys" to indicate the buoys where the information is available for more than 1000 instances. Furthermore, Hu *et al.* [34] used XGBoost and long short-term memory (LSTM) to obtain predictions with respect to WaveWatch III (WWIII) at a fraction of the computational cost at Lake Erie. Lou *et al.* [35] similarly proposed an LSTM-based network to forecast ocean wave heights using wave heights, wind speed and wind direction as input. A different line of study [36] used the moth-flame optimizer to fine-tune the hyperparameters of a neural network to obtain accurate short-term predictions.

Literature review suggests that ML methods have been previously used for data assimilation to obtain bias-corrected forecasts. Specifically, Deshmukh *et al.* [37] used a nested SWAN model with WWIII to forecast the residuals of the numerical predictions using ANNs for a wave rider buoy located off Puducherry, India. The study explored significant wave heights and dominant wave periods to show that numerical methods combined with ANNs give sustained performance over a longer forecast horizon. Londhe *et al.* [32] similarly used ANNs to correct the numerical residuals from Indian National Centre for Ocean Information Services. Likewise, Zhang *et al.* [38] used Gaussian process regression to incrementally predict the residuals of wave heights obtained from the SWAN model with the ground observations. While the prediction error decreases dramatically using residual correction as the forecast horizon increases, none of the previous studies used external predictors to improve wave forecasts [38]. Zhang *et al.* [38] showed that the wave height used from hindcasts can be used to improve prediction accuracy but they did not explore other variables such as winds and currents. Similarly, Mooneyham *et al.* [39] used a residual CNN-based network using spectral features from the buoys and the wave hindcasts as input to perform data assimilation for up to 24 h. However, they did not utilize other environmental features or take latitude and longitude into account, and the size of their data set was limited (three buoys). To alleviate the shortcomings of the previous methods, we propose a novel methodology to forecast significant wave heights in oceans. Taking the features from buoys, geographical location, and Ifremer hindcasts [40], we train a Transformer model [41] to forecast first-order differenced values of significant wave heights. While we use buoy and wind features as previously used in the literature, we have introduced other features from hindcasts like currents, sea–air energy flux, and directional spreading to predict significant wave heights.

The main contributions of this article are as follows.

1) We obtain highly accurate forecasts of significant wave heights in oceanic waters, outperforming numerical and ML approaches.
2) We use various features from buoys, wind/wave hindcasts, and geographical location to improve wave forecasts.
3) We forecast significant wave heights after calculating residuals from WWIII, which helps us obtain a physically consistent model.

TABLE I
STATISTICAL PROPERTIES OF SIGNIFICANT WAVE HEIGHTS

Variable	Min	Max	Average	Std
Significant wave heights (m)	0.02	15.42	0.99	0.50

Std refers to the standard deviation. The statistical properties of significant wave heights are calculated over the years 2010–2016 for all 92 buoys.

The rest of the article is organized as follows. Section II describes the problem formulation. Section III describes the experimental setup, i.e., WWIII model, data set properties, stationarity, training procedure, evaluation metrics, and the proposed architecture. Section IV shows the comparison of numerical forecasts and hindcasts and the performance of the Transformer model with state-of-the-art ML and numerical methods. Section V discusses the performance of the proposed model in light of the state-of-the-art approaches. Finally, Section VI concludes this article.

II. PROBLEM FORMULATION

The modeling of significant wave height can be formulated as in the following equation:

$$H_s[n+k] = \bar{H}_s[n+k] + f(x,n) \tag{1}$$

where k, x, and n represent the forecast horizon, environment variables, and temporal position at individual buoys, respectively. Similarly, $H_s[k]$ and $\bar{H}_s[k]$ represent the observed and predicted significant wave heights at step k, respectively. The function $f(x,n)$ thus represents the residuals of the first derivative, which is then modeled by the Transformer using the buoy data. Error-correcting problems in which the ML algorithms predict the innovation (divergence from the true state) are easier to solve, resulting in smaller ML models which require less training data to train [42], [43].

In terms of FFT components, the significant wave height H_s can be calculated as in the following equation:

$$H_s = 4\sqrt{m_o} \tag{2}$$

where m_o represents the variance of the FFT spectrum and is defined as in the following equation:

$$m_o = \int_{-\infty}^{\infty} S(f)df \tag{3}$$

where f and $S(f)$ refer to the frequency band and energy spectral density at the given frequency band, respectively, and after calculating the wave heights and several other bulk parameters, National Oceanic and Atmospheric Administration (NOAA) buoys store them for future use, which is utilized in this study [44].

The statistical properties of significant wave heights (H_s) are displayed in Table I.

III. Experimental Setup

A. WaveWatch III Model

The WWIII model obtained from Ifremer is implemented over the gridded bathymetry for the whole globe [40], [45]. The global grid system covers longitudes from 0° to 360° E and latitudes from 77° S to 77° N with a resolution of $0.5^\circ \times 0.5^\circ$. In WWIII, the evolution of the wavefield is obtained by solving the spectral wave balance equation by using the directional wave spectrum resolved at the grid points across the wave number direction bands, as in (5):

$$\frac{\partial N}{\partial t} + \frac{1}{\cos\phi}\frac{\partial}{\partial \phi}\dot{\phi} N \cos\theta + \frac{\partial}{\partial \lambda}\dot{\lambda} N + \frac{\partial}{\partial k}\dot{k} N + \frac{\partial}{\partial \theta}\dot{\theta}_g N = \frac{S}{\sigma} \tag{4}$$

where t, σ, S, N, and R refer to time, intrinsic angular frequency, sourcing term, action density, and radius of the earth, respectively. The terms $\dot{\phi}$, $\dot{\lambda}$, and $\dot{\theta}$ can be obtained using (6)–(8):

$$\dot{\phi} = \frac{c_g \cos\theta + U_\phi}{R} \tag{5}$$

$$\dot{\lambda} = \frac{c_g \sin\theta + U_\lambda}{R\cos\phi} \tag{6}$$

$$\dot{\theta}_g = \dot{\theta} - \frac{c_g \tan\phi \cos\theta}{R} \tag{7}$$

$$\dot{k} = -\frac{\partial \sigma}{\partial d}\frac{\partial d}{\partial s} - k\frac{\partial U}{\partial s} \tag{8}$$

where c_g, λ, ϕ, θ, k, U, and d refer to group speed of waves, longitude, latitude, wave propagation direction, wave number, wind speed, and depth, respectively. U_ϕ and U_λ refer to the current components in ϕ and λ directions, respectively.

We use the following source terms in the WWIII model used in this study: F90 NOGRB SCRIP SCRIPNC NC4 TRKNC DIST MPI PR3 UQ FLX0 LN1 ST4 STAB0 NL1 BT4 DB1 MLIM TR0 BS0 IC2 IS2 REF1 IG1 XX0 WNT2 WNX1 RWND CRT1 CRX1 TIDE O0 O1 O2 O2a O2b O2c O3 O4 O5 O6 O7. Other parameters used can be obtained from Ifremer [45].

WWIII is a numerical model based on the spectral representation of the sea state, i.e., at the selected locations, the wave field is decomposed into a spectrum, which defines the energy of many wave components. The sum of the wave trains going in different directions θ and with different frequencies f gives the whole sea state at a given location. Equation (5) can be solved forward to obtain a spectral representation of the ocean state. Afterward, (3) can be used to calculate the variance of the spectra, and (2) can be used to calculate the forecasted significant wave height. The forecasted significant wave height and the current wave height are then used in (1) to calculate the residuals. Finally, the Transformer algorithm is used to model the residuals using the environmental variables as features.

B. Transformers for Differential Equations

A Transformer is a deep learning method based on an encoder–decoder architecture that uses a self-attention mechanism [41]. Each layer in the encoder consists of a self-atten block and a feedforward network block. Both blocks implement a residual connection and a layer normalization unit [46].

Each Transformer block can be represented as

$$y_{t+1} = y_t + G(LN(y_t), \theta_t) \tag{9}$$

where $LN(.)$ is the layer normalization function and $G(.)$ is either a self-attention or feedforward layer [46]. y_t and θ_t refer to the output and features at position t, respectively. The iterative updates can be interpreted as discretization of continuous function transformations.

$G(LN(y_t), \theta_t)$ can be represented as function $F(y_t, \theta_t)$ for simplicity.

Afterward, if we relax y_t and θ_t to continuous functions $y(t)$ and $\theta(t)$, we can rewrite (9) as (10)

$$y(t + \Delta t) = y(t) + \Delta t \dot{F}(y(t), \theta(t)) \tag{10}$$

where Δt is the change of *t*, which is also called the step size. We can adjust Δt using a limit so that we get the following equation:

$$\lim_{\Delta t \to 0} \frac{y(t + \Delta t) - y(t)}{\Delta t} = F(y(t), \theta(t)). \tag{11}$$

From (13), we can deduce that each Transformer block describes a first-order differential equation [46], [47]. Moreover, based on the formulation of Transformer block in (9)–(11) and the universal approximation theorem of neural networks [48]–[51], complexity of the network can be increased to model higher order differential equations, given enough training data (e.g., increasing the number of encoder–decoder blocks, feedforward layers, hidden layers, etc.). Since $f(x, n)$ is a discrete version of $F(y(t), \theta(t)$, the Transformer model can be used to model the function $f(x, n)$ as in (1).

The Transformer architecture used in this article is illustrated in Fig. 1.

Fig. 1 shows the Transformer model used in this study. The model includes decoder and encoder blocks, with each containing an embedding layer, multihead self-attention layer, and, finally, feedforward layers in each block. The primary purpose of the embedding layer is to obtain an $n = 5$ dimensional expanded representation of the markers. A feedforward neural network is used as the embedding layer. The input of the embedding layer is m markers, and the output is an $m \times n$ vector. The output of the embedding layer is then passed to the multihead attention layer. The positional encoding of features allows the network to learn various cycles in the data. Similarly, the attention mechanism learns the correlation between the features at each step of the sequence, filters out unnecessary information, and assigns higher weights to the important features, thus allowing the network to learn long-term dependencies and give better performance for time-series data. Transformer networks are generally used for natural language processing tasks in which sentences/words are encoded using byte-pair encoding and then passed to the input layer. However, since we have numerical features derived from buoys, hindcast, and geographical coordinates, we do not need to obtain byte-pair encoding. Therefore, we use StandardScaler [52] to scale the features and pass them directly to the input layer of the Transformer model.

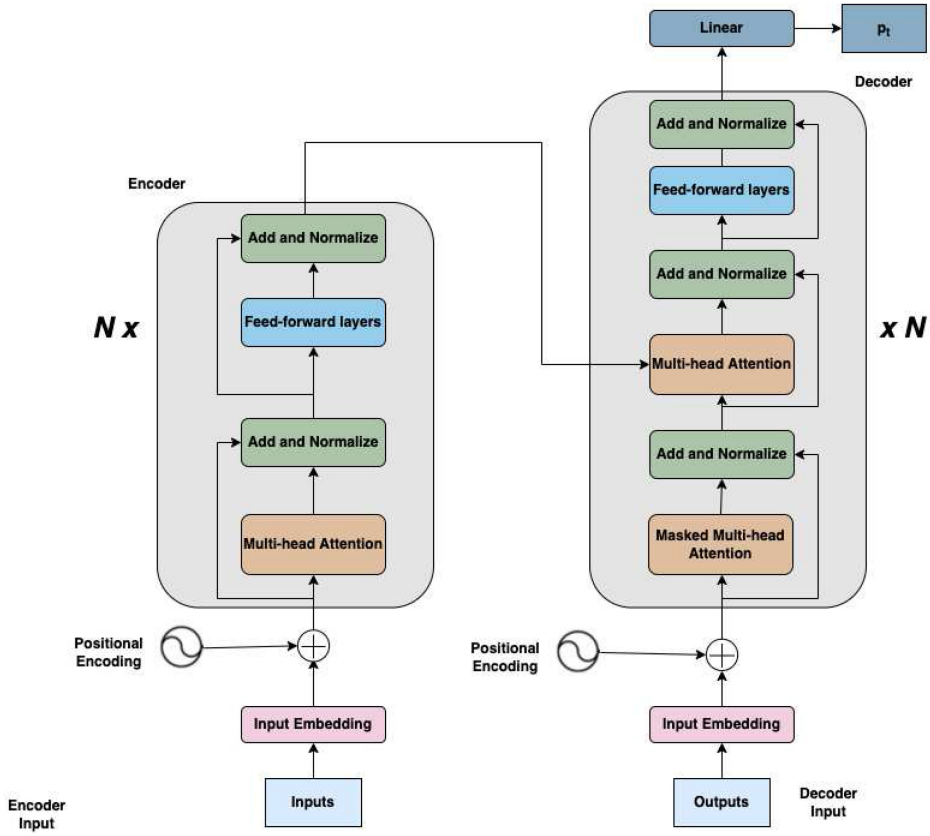


Fig. 1. Architecture of the Transformer network used in this article. Left side denotes encoder block and right side denotes decoder block.

The multihead attention layer used in this article is based on the self-attention mechanism. The input of this layer is the expanded representations of the inputs obtained from the embedding layer. The self-attention mechanism then calculates the attention score for all other expanded representations of inputs. The attention score can be calculated as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (12)$$

where Q, K, V, and d_k refer to the matrix of all the queries, matrix of all the keys, matrix of all the values, and the key dimension, respectively. In a multihead attention setting, the Transformer model creates h independent linear representations from queries, keys, and values. These h representations are normalized and then passed to the linear projection layer to obtain the final output from the encoder/decoder layer. The term N in Fig. 1 shows that multiple encoder/decoder blocks can be used to create a Transformer neural network (TNN). However, we use a multihead attention layer with two heads in each encoder/decoder layer for our setup. Similarly, the network contains four layers with 200 neurons for each feedforward layer in each encoder/decoder block. The dimensionality of input and output is $d_{\text{model}} = 512$ and the inner layer dimensionality is $d_{ff} = 2048$.

C. Data Set

We use 92 NOAA buoys to compare various numerical methods. The predictions and WWIII forecasts for these buoys can be obtained from Ifremer hindcast [40]. The data is obtained in various formats like time-series, 30 min averaged, and spectral data. To compare with the state-of-the-art ML methods, we use a limited number of buoys, which contain the environmental parameters used by other ML methods. The list of buoys is provided in the Appendix section (Part D).

The features used in this study include the following.

1) Buoy features: The buoy features used in this study are significant wave heights, mean wave period, dominant wave periods, wave direction, kurtosis, and power spectral density.
2) Geographical location: To quantify the geographical location of the buoys, we use latitude and longitude as variables.
3) Hindcast features: The hindcast features used in this study include sea–air energy flux, U and V components of currents, U and V components of winds, and directional spreading.

D. Stationary Property of Data

Nonstationary data, by definition, is unpredictable and cannot be modeled or forecasted [53]. The results that are obtained

TABLE II
ADF AND KPSS STATISTICS ON THE BUOY WITH NOAA IDENTIFIER 41010

Points	ADF		KPSS	
	t-Statistics	*p*-Value	*t*-Statistics	*p*-Value
100	−1.8510	0.3552	0.1950	0.100
500	−3.8590	0.002	0.1379	0.100
1000	−5.2011	8.74×10^{-6}	0.2973	0.100
5000	−8.9326	9.72×10^{-15}	0.3467	0.100

Stationarity of the series with first-order differencing performed with the data from the years 2013, 2014, and 2015.

TABLE III
ADF AND KPSS STATISTICS ON THE BUOY WITH NOAA IDENTIFIER 51101

Points	ADF		KPSS	
	t-Statistics	*p*-Value	*t*-Statistics	*p*-Value
100	−4.0434	0.003	0.2176	0.1000
500	−5.5625	2.4×10^{-10}	0.1167	0.10000
1000	−9.9887	4.87×10^{-17}	0.0548	0.10000
5000	−16.6592	1.58×10^{-29}	0.0756	0.10000

Stationarity of the series with first-order differencing performed with the data from the years 2013, 2014, and 2015.

by fitting a model in nonstationary time series may be spurious and may indicate a relationship between variables where none exists [53], [54]. To measure the stationary property of the data, tests like augmented Dickey–Fuller (ADF) [55] and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) [56] have been proposed.

ADF and KPSS complement each other's strengths in determining if a series is stationary by proving the existence or absence of a unit root. A unit root denotes the stochastic trend (random walk with a drift), which shows a systematic pattern that is not predictable. Therefore, the rejection of the null hypothesis of ADF by the data is an indication that the series has no unit root. However, in the case of KPSS, rejecting the null hypothesis means that the series has a unit root. KPSS and ADF tests are performed on the data from the buoy with NOAA identifier 41010 after the first-order differencing. These two tests are also performed on the buoy with NOAA identifier 51001. Both buoys were selected randomly.

For 100 points, Table II shows that the data rejects the ADF test but not the KPSS test. However, as the number of points increases, the *p*-value for ADF of the data points goes down from 0.3552 at 100 points to 9.72×10^{-15} at 5000 points. A similar trend is exhibited by buoy 155, as shown in Table IV.

The *p*-values from Tables II and III show that as the number of points increases, the stationary trend can be found through both ADF and KPSS tests. From our test on two buoys (NOAA identifiers 41010 and 51101), the minimum number of points used should be around 500 for each buoy. Thus, the availability of sufficient data at different times and geographical locations should enable the algorithms to learn the trend and provide meaningful and accurate forecasts.

TABLE IV
EVALUATION METRICS AND THEIR CALCULATIONS

Name	Mathematical formula
MSE	$\frac{\sum_{i=1}^{N}(x_i-\bar{x}_i)^2}{N}$
RMSE	$\sqrt{\frac{\sum_{i=1}^{N}(x_i-\bar{x}_i)^2}{N}}$
MAE	$\frac{1}{N}\sum_{i=1}^{N}\lvert x_i-\bar{x}_i\rvert$
SI	$\frac{\text{RMSE}}{\frac{1}{n}\sum_{i=1}^{N}x_i}$
CC	$\frac{\sum_{i=1}^{N}(x_i-x_m)(\bar{x}_i-\bar{x}_m)}{\sqrt{\sum_{i=1}^{N}(x_i-x_m)^2\sum_{i=1}^{N}(\bar{x}_i-\bar{x}_m)^2}}$
Bias	$\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x}_i)$

In the preceding table, i, x_i, $\bar{x}_i$, x_m, $\bar{x}_m$, and N refer to the position, measured value, predicted value, mean of actual values, mean of predicted values, and the number of elements, respectively.

E. Evaluation Metrics

To measure the performance of our model and to compare the results with state-of-the-art numerical and ML methods, we employ various metrics like mean square error (mse), root mean square error (RMSE), mean absolute error (MAE), variance, R^2 score, scatter index (SI), Pearson's correlation coefficient (CC), hanna and heinold indicator (HH), and bias. Note that RMSE, bias, and MAE are measured in meters, and SI, R^2 score, HH, and CC are nondimensional. While we use MSE as the cost function for optimizing the Transformer model, we use the other metrics to compare with other methods. The metrics used are displayed in Table IV.

F. Proposed Methodology

In this article, we propose a Transformer [41] based framework to forecast the significant wave heights as displayed in (1). We then use a sliding window of a size equal to the forecast horizon. The features are then fed to the Transformer, where the forecasting is performed. After the Transformer model performs the forecasting, the predictions from WWIII are added to the forecasted values to generate final predictions.

Fig. 2 shows the proposed methodology for forecasting significant wave heights in oceanic waters. The features used in this study are spectral density, kurtosis, latitude, and longitude. First, the Transformer model is used to model the residuals after first-order differencing. Afterward, the prediction of the residuals from the Transformer model is fed to (1) to forecast the significant wave heights.

Transformer architectures employ a self-attention mechanism to learn relationships between the elements of a sequence [57] which makes it suitable for our study. Self-attention is also

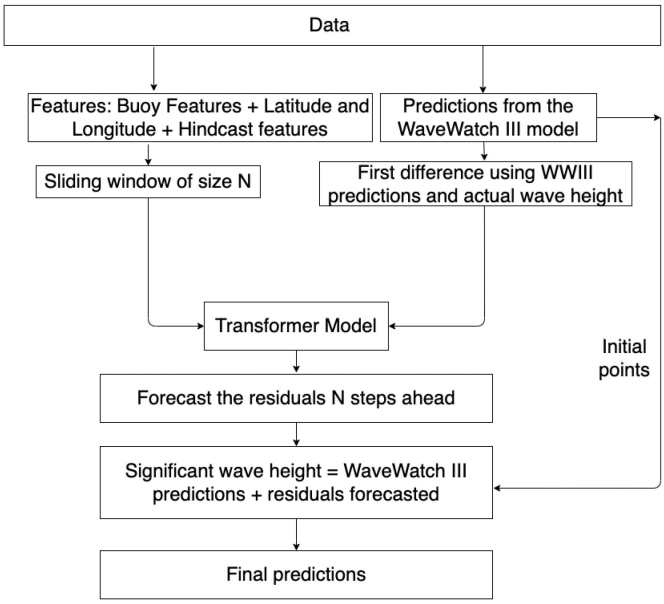


Fig. 2. Proposed methodology. The figure shows the proposed framework for forecasting wave heights.

invariant to permutations and changes in the number of input points, allowing it to operate on irregular grids. The self-attention mechanism, which captures long-term dependencies in the data, motivates our study's use of the Transformer model.

IV. Results

In this section, Section IV-A and IV-B show the comparison of the proposed framework with existing ML and numerical methods, respectively.

A. Comparison of Numerical Forecasts and Hindcasts

Owing to the lack of the availability of long-term reforecast data to train the ML models, we use the hindcasts from WWIII as the proxy to numerical forecasts following the procedure used by Mooneyham *et al.* [39]. The authors report that since the WWIII hindcasts are driven using reanalysis winds, they have higher prediction skill than operational models run without ground-truth observations. The higher skilled proxy serves as a comparatively difficult benchmark for ML algorithms leaving little room for improvement.

First, we report the results from Bidlot [58] for the operational forecasts of different forecast agencies. Note that the forecasts are operational; so changes might be made to the files in real time as errors are discovered, thus leaving different agencies with different forecasts even if the setup might be the same.

1) Prediction Skill of Various Operational Forecasts: We first investigate the prediction error of various weather agencies for operational forecasts. The operational forecasts are collected for the numerical forecasts at the overlapping buoys between different forecast agencies. Then the results are accumulated monthly and yearly by Bidlot *et al.* [58]. The results for the year 2016 are displayed in Table V.

TABLE V
Comparison of Operational Forecasts for Hindcasting and 48 H Ahead Forecasting

Ref.	Lead time (h)	RMSE (m)	BIAS (m)	CC	SI
ECMWF	0	0.278	−0.001	0.979	0.131
	48	0.356	−0.022	0.965	0.168
MO	0	0.328	0.053	0.972	0.153
	48	0.428	0.079	0.954	0.199
FNMOC	0	0.279	−0.006	0.979	0.131
	48	0.457	0.043	0.948	0.215
NCEP	0	0.339	−0.084	0.970	0.155
	48	0.408	−0.014	0.954	0.193
MF	0	0.291	−0.084	0.977	0.137
	48	0.377	−0.064	0.962	0.176
DWD	0	0.307	0.036	0.979	0.144
	48	0.433	0.034	0.948	0.204
BoM	0	0.345	−0.110	0.970	0.154
	48	0.431	−0.066	0.951	0.202
SHOM	0	0.335	−0.069	0.972	0.155
	48	0.399	−0.065	0.957	0.186

Prediction skill of operational forecasts from various weather agencies using different WWIII settings. The operational forecasts were obtained by Bidlot *et al.* [58] and stored for reference. Since the operational models change over time, the specific configuration used is not available.

Since the report of Bidlot *et al.* [58] did not consider other forecast horizons, we do not display them in this article. The models compared are European center for medium-range weather forecasts (ECMWF) [59], [60], Met Office (MO) [61], Fleet Numerical Meteorology and Oceanography Centre (FNMOC) [62]–[64], Meteorological Service of Canada [65], [66], National Centers for Environmental Prediction (NCEP) [67], [68], Meteo France (MF) [69], [70], Deutscher Wetterdienst (DWD) [71], Bureau of Meteorology (BoM) [59], [72]–[74], Service Hydrographique et Oceanographique de la Marine (SHOM) [75], Japan Meteorological Agency [76], and Korea Meteorological Administration [77]. The reader is referred to Bidlot *et al.* [59] for buoys information.

Table V shows that the prediction accuracy of the numerical models does not decrease rapidly up to about 48 h ahead. The highest difference is the FNMOC model with the RMSE of 0.339 m for nowcasting and 0.457 m for 48 h ahead of forecasting. For the NCEP model-driven using climate forecast system reanalysis (CFSR) winds [78], [79], the WWIII model has the RMSE of 0.339 m for nowcasting and 0.408 m for 48 h ahead forecasting. Note that the RMSE of the hindcast on the data set is 0.356 m. The decreased RMSE of hindcast compared to the operational forecasts suggests that hindcasts have relatively higher skills than operational forecasts. However, there remains room for correction of error.

2) Comparison Between Wavefields Obtained Using WaveWatch III and Forecasts: Mooneyham *et al.* [39] have previously used a spectral net-based setup to perform data assimilation using WWIII hindcasts and a convolutional residual network. The authors suggested that reanalysis driven hindcasts have higher prediction skills than the operational

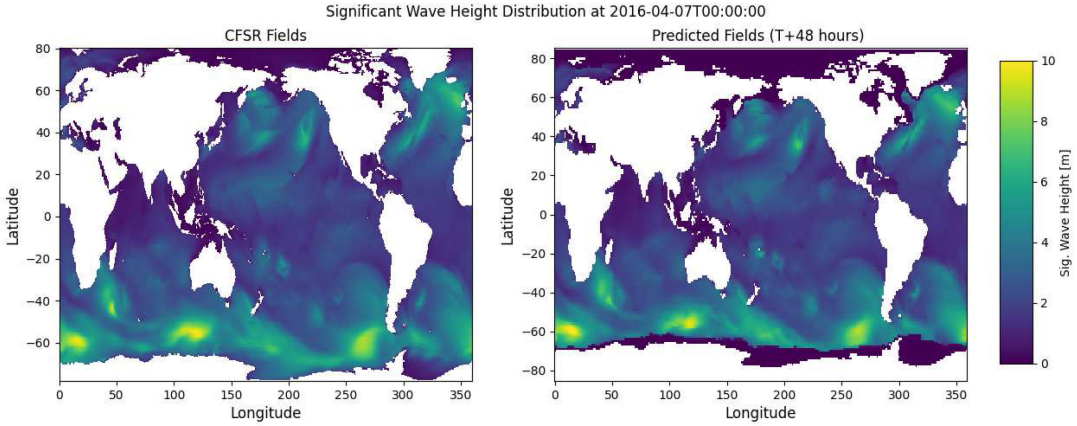


Fig. 3. Comparison between wavefields obtained from the CFSR hindcasts (left) and the forecasts (right) obtained using the WWIII model without the forcing applied for the forecast horizon.

forecasts driven using analysis fields, thus providing a difficult benchmark to make improvements to WWIII predictions using ML methods. However, to make reliable assumptions, we perform experiments with a single grid coarse resolution WWIII model to show that the wave fields persist up to 48 h, which is the most extended forecast horizon for our experiment setup.

We first investigate the difference in forecast skills between the hindcast wave fields and the forecast wave field obtained by running the model forward in time without the wind forcing. Then, we first spin up the WWIII model for 30 days for the forecast. We use a coarse resolution global grid while the hindcast data is obtained from a high-resolution multigrid WWIII run. This choice is made due to the high computational complexity of running the numerical model for high-resolution grids for a longer period.

The global grid system covers longitudes from 0° to 360° E and latitudes from 85° S to 85° N with a resolution of $1^\circ \times 1^\circ$. Time-stepping for the model occurs with a global time step of 30 min, with sub-steps in the fractional step integration sufficiently small to ensure model stability and accuracy (specifically 450 s for spatial advection, 900 s for intraspectral propagation, and a minimum dynamic source time step of 10 s). Similarly, the frequency-direction space is discretized over the full circle with 36 directions and 36 frequencies. The directional grid has a constant spacing of 10°, whereas the frequency grid is logarithmically distributed with growfactor 1.1, starting from $f_1 = 0.035$ Hz and ending at $f_3 = 0.98$ Hz. Three hourly analysis fields from the CFSR wind archive [78], [79] and weekly ice concentration from NCEP [80] are used to force the model. The hindcast is driven by CFSR winds [78], [79], daily ice concentration, and sea surface temperature from the passive microwave radar fields. Since the ice concentration used in the coarse resolution single grid forecasting setup is weekly instead of daily as in the hindcast archive, some differences are expected. The comparison between CFSR hindcast fields and the WWIII model is displayed in Fig. 3.

From Fig. 3, we can see that while the forecast model used overestimates the high-intensity wave fields in the Pacific Ocean, wave height distribution from both CFSR hindcasts and the WWIII model used in this study for forecasting look similar. The overestimation can be attributed to the low-resolution grids and the weekly ice concentration fields used in this study. The forecast wavefields are obtained from a single global grid of coarse resolution of $1^\circ \times 1^\circ$, while the hindcasts are obtained using the multigrid WWIII setup with the global grid resolution of $0.5^\circ \times 0.5^\circ$. Moreover, we can see that the model slightly underestimates the wave fields in the Antarctic ocean, attributing to the weekly ice concentration fields.

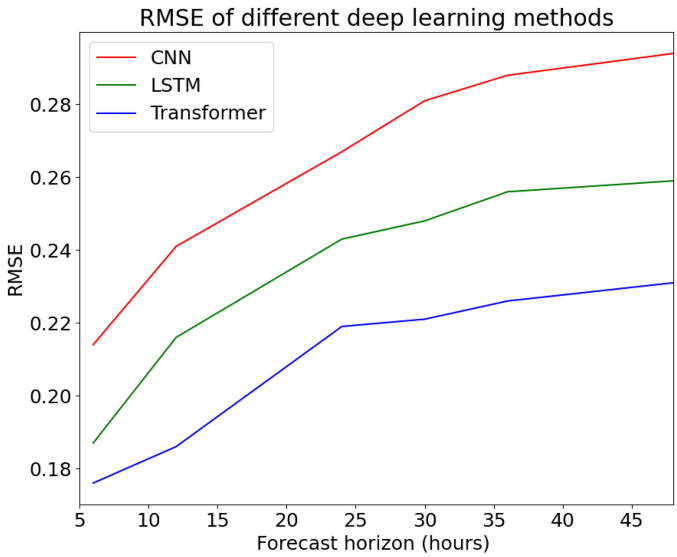


Fig. 4. RMSE values of CNN, LSTM, and Transformer for different forecast horizons using WaveWatch III hindcasts.

B. Experiments With Features and Different Differencing Schemes

In this subsection, we measure the effect of various features and the differencing schemes to obtain the best settings for ML correction of numerical wave heights. We first display the

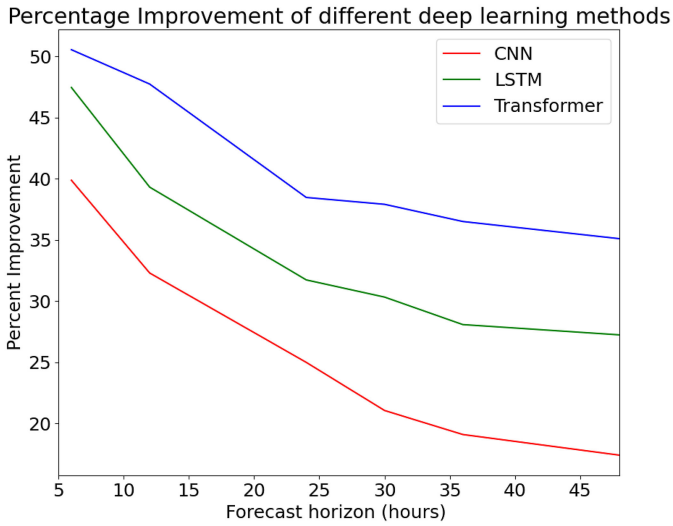


Fig. 5. Percentage improvements of CNN, LSTM, and Transformer for different forecast horizons using WaveWatch III hindcasts.

TABLE VI
COMPARISON OF TRANSFORMER NEURAL NETWORK WITH VARIOUS FEATURE COMBINATIONS

Features	Lead time (h)	RMSE (m)
Buoy features	6	0.971
	12	0.980
	24	0.983
	30	0.985
	36	0.987
	48	0.992
Buoy features + latitude/longitude	6	0.291
	12	0.465
	24	0.646
	30	0.689
	36	0.714
	48	0.764
All features	**6**	**0.241**
	12	**0.365**
	24	**0.458**
	30	**0.462**
	36	**0.471**
	48	**0.487**

Performance of the Transformer neural network with various feature combinations. **Bold** represents the setup with the best performance. Note that no differencing has been applied.

performance of the TNN with various features and select the best combination in Table VI.

Table VI shows that for the setup with no differencing, the performance of the TNN increases as we increase the number of features. Previous studies can explain the increase in prediction skill with deep neural networks, suggesting that they automatically perform feature selection/extraction [81]. We, thus, use all the features available for the subsequent predictions. The final features contain buoy features (wave heights, power spectral density, kurtosis, mean wave periods, dominant wave periods, and wave direction), geographical features (latitude and longitude), and other features derived from the hindcasts (sea–air energy flux, U and V components of currents, U and V components of winds, and directional spreading) [40].

TABLE VII
COMPARISON OF TRANSFORMER NEURAL NETWORK WITH VARIOUS DIFFERENCING SCHEMES

Differencing order	Lead time (h)	RMSE (m)
No differencing	6	0.241
	12	0.365
	24	0.458
	30	0.462
	36	0.471
	48	0.487
First-order differencing	**6**	**0.172**
	12	**0.215**
	24	**0.228**
	30	**0.242**
	36	**0.245**
	48	**0.249**
Second-order differencing	6	0.212
	12	0.245
	24	0.274
	30	0.293
	36	0.301
	48	0.304

Performance of the Transformer neural network with various differencing schemes. **Bold** represents the setup with the best performance.

After obtaining the best set of features, we now test various differencing schemes with the same set of features. We perform first-order differencing by using the equation $H_s[n+k] - \bar{H}_s[n+k]$. The second-order differencing is subsequently performed by subtracting $\bar{H}_s[n+k] - \bar{H}_s[n]$ from the first-order difference, i.e.,

$$(H_s[n+k] - \bar{H}_s[n+k]) - (\bar{H}_s[n+k] - \bar{H}_s[n]) = H_s[n+k] - 2\bar{H}_s[n+k] + \bar{H}_s[n]. \quad (13)$$

The differencing scheme used to obtain second-order difference is known as variable step size differencing [82]. The performance of the Transformer model with various differencing schemes is displayed in Table VII.

Table VII shows the performance of the TNN with various differencing features using all the available features. We do not utilize the sliding window scheme while testing various differencing orders. Since the first-order differenced system obtains the best performance among all the differencing setups, we choose the first-order setup for the rest of the experiments.

The Transformer model without differencing obtains an RMSE of 0.241 m for 6 h ahead prediction and 0.487 m for 48 h ahead prediction. Meanwhile, the first-order differenced scheme obtains an RMSE of 0.172 m for 6 h ahead prediction and 0.249 m for 48 h ahead prediction. The difference in prediction

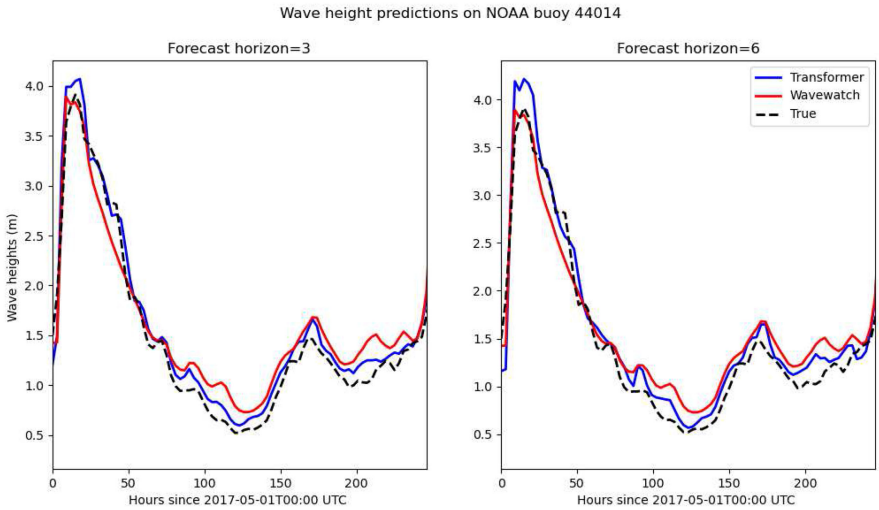


Fig. 6. Wave height plots of WaveWatch III, Transformer network, and the ground-truth values for NOAA buoy 44014 for 3 and 6 h ahead forecast horizons.

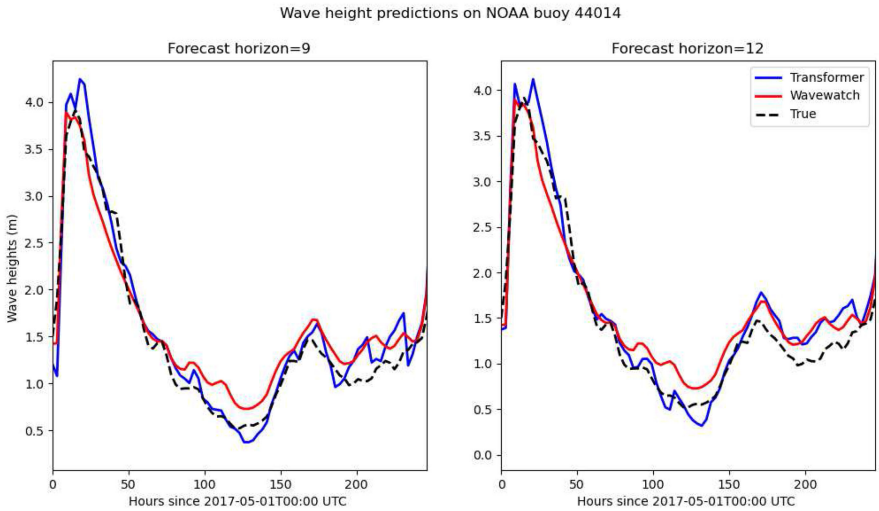


Fig. 7. Wave height plots of WaveWatch III, Transformer network, and the ground-truth values for NOAA buoy 44014 for 9 and 12 h ahead forecast horizons.

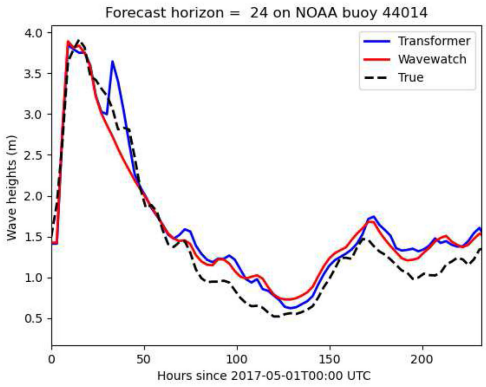


Fig. 8. Wave height plots of WaveWatch III, Transformer network, and the ground-truth values for NOAA buoy 44014 for 24 h ahead forecast horizon.

skill suggests that while the Transformer model without the differencing underperforms the operational numerical methods shown in Tables V–VII, it can still help correct wave parameters from the numerical model since, unlike the traditional data assimilation methods, it can easily take external predictors into account.

C. Comparison With State-of-the-Art Machine Learning Approaches

In this subsection, we compare the performance of the proposed Transformer model with the state-of-the-art ML approaches for significant wave heights forecasting. Since Li and Liu [29] perform the experiments on the data collected from NOAA buoy 41002 (February 1–16), we perform the

TABLE VIII
PERFORMANCE OF THE PROPOSED ALGORITHM WITH DIFFERENT MACHINE LEARNING METHODS FOR SHORT-TERM FORECASTING

Ref.	Lead time (h)	CC	SI	Model	Feature
Ozger	3	0.960	0.131	Fuzzy-logic	U_{10}
[83]	6	0.899	0.211		H_s
	12	0.800	0.289		
Ozger	3	0.925	0.184	ARMAX	U_{10}
[84]	6	0.842	0.260		H_s
	12	0.690	0.349		
[20]	1	0.986	0.014	WD-SVR	H_s
	3	0.954	0.044		
	6	0.855	0.086		
[30]	3	0.950	0.211	RF	H_s
	6	0.910	0.279		U_{10}, W_t
	12	0.831	0.607		ω, A_t
	24	0.734	0.748		Pd
[29]	3	0.984	0.049	Bayesian network	Gst
	6	0.933	0.063		U_{10}
	12	0.914	0.073		T_a
	24	0.752	0.168		A_t
Proposed method	3	**0.999**	**0.027**	**Transformer**	$\boldsymbol{Buoy}$
	6	**0.989**	**0.037**		$\boldsymbol{Lat/Lon}$
	12	**0.948**	**0.054**		$\boldsymbol{Hindcasts}$
	24	**0.935**	**0.062**		

Bold indicates the best value. Likewise, U_{10}, H_s, T_a F_v, D_{mean}, W_t, A_t, ω, Gst, and Pd refer to wind speed, significant wave heights, average wave period, friction velocity, wave direction, water temperature, air temperature, pressure, gust speed, and dew point, respectively. Similarly, $\boldsymbol{Buoy}$, $\boldsymbol{Lat/Lon}$, and $\boldsymbol{Hindcasts}$ refer to buoy features (wave heights, dominant wave periods, power spectral density, and kurtosis), geographical features (latitude and longitude), and the features derived from hindcasts (sea–air energy flux, U and V components of currents, U and V components of winds, and directional spreading). The other state-of-the-art methods are only executed up to the time horizons they were originally proposed.

experiments on the same data set. The other state-of-the-art methods have only been executed up to the time horizons they were originally proposed [29]. The comparative performance is displayed in Table VIII. The methods compared with are those of Duan *et al.* [20], Mafi *et al.* [30], and Liu [29]. Further, the fuzzy-logic [83] and ARMAX [84] based methods proposed by Ozger [23] are also included in the comparative analysis. The Transformer model used in this subsection uses a sliding window of size equal to the forecast horizon.

The proposed method with the Transformer model outperforms other methods with an SI of 0.062 m for 24 h ahead prediction compared to 0.168 m for [29], and 0.748 m for Mafi *et al.* [30].

D. Comparison With Other Deep Learning Approaches Using Numerical Residuals

This subsection displays various deep learning approaches on the numerical residuals. Following the previous literature on the topic, we compare convolutional neural networks (CNNs) and LSTMs with the proposed method. Note that previous studies have already used CNN and LSTMs for significant wave height forecasting [34], [35], [39]. All deep learning models used in this subsection use sliding windows equal to the forecast horizon. In the case of CNN, a Conv2D layer is used from PyTorch with four hidden layers with 200 neurons each to make predictions. The network afterward consists of four hidden layers with 200 neurons each. Similarly, for LSTMs, we use two bidirectional LSTM layers and then use four hidden layers with 200 neurons each afterward. All other parameters are kept constant. The train:test:validation split is set to (0.6:0.3:0.1).

From Fig. 4, we can see that the RMSE of all deep learning methods gradually increases as the forecast horizon increases. Specifically, in the case of CNN, the RMSE values are 0.214, 0.241, 0.267, 0.281, 0.288, and 0.294 m for 6, 12, 24, 30, 26, and 48 h, respectively. Likewise, the LSTM model obtains RMSE values of 0.187, 0.216, 0.243, 0.248, 0.256, and 0.259 m for 6, 12, 24, 30, 26, and 48 h, respectively. Finally, the Transformer model obtains RMSE values of 0.176, 0.186, 0.219, 0.221, 0.226, and 0.231 m for 6, 12, 24, 30, 26, and 48 h, respectively.

Similarly, Fig. 5 shows that the percentage improvement on the residuals from WWIII decreases for all models as the time horizon increases. Specifically, CNN has the most negligible percentage improvement over the other two methods with percentage improvements of 39.9%, 32.3%, 25.0%, 21.1%, 19.1%, and 17.1% for 6, 12, 24, 30, 36, and 48 h, respectively. In the case of the other two methods, LSTM and Transformer, which use specialized structures to capture the temporal correlations, the improvement persists for longer. However, the transformer model outperforms the other models with percentage improvements of 50.5%, 47.8%, 38.5%, 37.9%, 36.2%, and 35.1% for 6, 12, 24, 30, 36, and 48 h, respectively. This is higher than percentage improvements obtained from the LSTM model, which are 47.5%, 39.3%, 21.7%, 30.3%, 28.1%, and 27.3% for 6, 12, 24, 30, 36, and 48 h, respectively.

The Transformer model employs a multihead attention layer and encoder–decoder blocks to learn long-term dependencies in the data [41], [57]. Figs. 4 and 5 show that the CNN model, which does not have any structure to take temporal dependencies into account, performs the worst. Moreover, the LSTM model, which can capture some temporal dependencies, obtains performance that is better than CNN but worse than TNN.

V. DISCUSSION

The proposed methodology outperforms the existing numerical and ML approaches, as displayed in Tables V–VIII. However, the numerical schemes that solve differential equations with exact conditions do not generalize well to all physical conditions. Since the exact conditions make the schemes inflexible, the numerical solutions have to be reinitialized for a grid, and then the value propagated over the respective grid. Moreover, the updates to the numerical models are done iteratively, which accumulates the errors, and, thus, the prediction capability decreases as the forecast horizon increases. Since the proposed methodology uses the predictions from the WWIII model and then predicts the residuals to generate wave height predictions, the performance is better than the other numerical models.

The state-of-the-art ML methods compared do not take numerical predictions into account while making forecasts. Moreover, the complex physics arising from interactions between various environmental components is not considered. Figs. 6–8 show the difference in forecast skills between WWIII and corrections using TNN. To generate these plots, the Transformer

network is rerun with the same setup on the NOAA buoy 44014 for the period January 2010 to April 2015. We use the train:validation split of (0.7:0.3) to train the model and then plot the predictions of the resulting models in Figs. 6–8.

Figs. 6–8 display the predictions of WWIII hindcasts and the corrections after applying TNN for buoy with NOAA Identifier 44014 for May 2015. For Fig. 6, which shows the plot of 3 h ahead residual correction using Transformer network, the predictions are close to the true values. However, as the forecast horizon increases to 24 h, the prediction skill of TNN decreases, and the model overpredicts anomalous waves similar to the WWIII. The decrease in prediction skills suggests that while the statistical model helps improve short-term forecasts, numerical methods contribute the most when forecasting longer horizons.

While statistical models are considered accurate for nowcasting up to 6 h, they perform poorly while making predictions for more extended time horizons [85]. From Table VII, we can see that the performance of the TNN is better after first-differencing than without it. The removal of autocorrelations arising from the deterministic components provides the data on which ML methods can be used to make predictions. This procedure also decreases the generalization error significantly, as displayed in Section IV. After differencing, the generalization error decreases, suggesting that combining them yields better performance than using ML or numerical methods individually for forecasting significant wave heights. Moreover, since the prediction performance decreases slightly for second-order differencing, it suggests that some noise is inserted when the order of differencing increases above the order of 1.

The results of CNN used in this study, as shown in Figs. 4 and 5, are comparable to the recent work of Mooneyham *et al.* [39] who used a residual CNNs to obtain bias-corrected forecasting using only the spectral features as input. While the authors used a neural network layer with eight hidden layers, which is two times higher than the ones used in this study, the error-correction is similar, and the least skillful CNN model also retains forecast skill up to 48 h with 17.4% improvement compared to 10%–20% for Mooneyham *et al.* up to 24 h. Furthermore, the authors have 23%–50% error reduction for the first 6 h, which is comparable to ours. Compared to a similar setup, this improvement suggests that data and the analysis fields derived from the hindcasts are essential in data assimilation using neural networks.

The performance of LSTM and Transformer models, which retain their prediction skill even when the forecast horizon increases, suggests that the mechanisms used in those networks are essential in capturing time dependencies, which are missing from the numerical models. Moreover, since the Transformer model outperforms other setups, we can infer that the attention mechanism can capture long-range dependencies, thus yielding superior results.

VI. Conclusion

In this article, we have proposed a Transformer-based framework for highly accurate prediction of significant wave heights in oceans using buoy data. The proposed method significantly outperforms the state-of-the-art ML and numerical methods on a case study performed using WWIII hindcasts as a numerical proxy. The proposed methodology obtains the RMSE of 0.231 m for two days ahead forecasting.

We have used various features like wave heights, average wave period, dominant wave periods, wave direction, power spectral density, and kurtosis from the buoys. Similarly, we have used U and V components of wind, U and V components of currents, sea–air energy flux, and directional spreading, along with latitude and longitude for the prediction of significant wave heights. Since the proposed method uses QC procedures, numerical residuals, various relevant features, and the Transformer model that differentially weighs the input features, the performance of the proposed framework does not decrease significantly with the forecast horizon.

The Transformer is a deep learning method that uses an attention mechanism with an encoder–decoder architecture to perform predictions. The attention model differentially weighs the significance of each part of the input data, which helps better identify the context that confers meaning to various parts of a sequence. The Transformer model can thus be used to infer context from the lagged variables used as features in this study.

Wave forecasting can be performed using model equations of empirical relationships between various wave parameters. Compared to the nonlinear differential equations, ML methods provide similar models with lower computational complexity. While differencing helps make the data stationary, the Transformer model captures various nonlinear interactions to improve prediction performance. We posit that similar frameworks can be used to forecast other wave properties in oceans.

VI. Appendix

A. Derivation of Power Spectral Density and Kurtosis

Since the variance of the spectrum is calculated over all frequency bands, if we take the moments of individual frequencies, sample variance [86], [87] can be calculated as in the following equation:

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2 \tag{14}$$

where $\bar{X}$ is the sample mean, and N refers to the number of samples.

Now, the expectation of S^2 is defined as in the following equation:

$$E[S^2] = \sigma^2 \tag{15}$$

and the variance as in the following equation:

$$\text{Var}[S^2] = \frac{1}{N}\left(\mu_4 - \frac{N-3}{N-1}\sigma^4\right) \quad \text{for } N > 1 \tag{16}$$

where $\frac{\mu_4}{\sigma^4}$ is the kurtosis and σ^2 is the standard deviation.

Likewise, if the variables are uncorrelated, according to the Bienayme formula [88], the variance of random variables is equal to the sum of their variances as displayed in the following

equation:

$$\text{Var}\left(\sum_{i=1}^{N} X_i\right) = \sum_{i=1}^{N} \text{Var}(X_i). \tag{17}$$

There exists a similar property for kurtosis if it is represented in terms of excess kurtosis [89] which is displayed in the following equation:

$$\gamma = \frac{\mu_4}{\sigma^4} - 3 = \frac{1}{\sum_{j=1}^{N} \sigma_j^2} \sum_{i=1}^{N} \sigma_i^4 \gamma_i. \tag{18}$$

We, thus, use σ^4 and γ as the features in our study for forecasting significant wave heights. Variance is derived from the power spectral density and represents the energy of the system [90]. Likewise, the excess kurtosis can also be interpreted as a Benjamin–Feir index function [91] arising from the nonlinear Schrodinger equation in oceanic waters. Note that since Transformer is able to learn nonlinear relationships, we use the normalized versions of features for the study.

B. Calculation of Kurtosis from the Wave Spectra

Calculation of kurtosis is performed using Kuik *et al.* [92] estimate of kurtosis as in the following equation:

$$\begin{bmatrix} a_1 \\ b_1 \\ a_2 \\ b_2 \end{bmatrix} = \frac{1}{E^b} \int_{0.025}^{0.580} \left(df E_r(f) \begin{bmatrix} a_1(f) \\ b_1(f) \\ a_2(f) \\ b_2(f) \end{bmatrix} \right) \tag{19}$$

where a_1, a_2, b_1, and b_2 are the Fourier coefficients, and E^b is the variance. The variance is calculated using the following equation:

$$E^b = \int_{0.025}^{0.580} df E_r(f). \tag{20}$$

Afterward, we calculate m_1, θ, m_2, and kurtosis as in the following equations:

$$m_1 = \left(a_1^2 + b_1^2\right)^{\frac{1}{2}} \tag{21}$$

$$\theta = \tan^{-1}\left(\frac{b_1}{a_1}\right) \tag{22}$$

$$m_2 = a_2 \cos(2\theta) + b_2 \sin(2\theta) \tag{23}$$

$$\text{kurtosis} = \frac{\mu_4}{\sigma^4} = \gamma + 3 = \frac{6 - 8m_1 + 2m_2}{[2(1 - m_1)]^2} \tag{24}$$

where m_1, m_2, θ, and γ represent first-order moment, second-order moment, mean direction, and excess kurtosis, respectively.

C. Parameters of Transformer Neural Network

1) Optimizer: Adam (learning_rate = 0.0005, gamma = 0.95), epochs = 500, loss function = mean square error.

D. Buoys Used in the Study

1) Comparison With Numerical Methods: The 92 NOAA buoys used for comparison with numerical methods are:

41008, 41009, 41010, 41013, 41025, 41040, 41041, 41043, 41044, 41046, 41047, 41048, 41049, 42001, 42002, 42003, 42012, 42019, 42020, 42035, 42036, 42039, 42040, 42055, 42056, 42057, 42059, 42060, 44007, 44009, 44013, 44014, 44020, 44025, 44027, 44065, 44066, 45001, 45002, 45003, 45004, 45005, 45006, 45007, 45008, 45012, 46001, 46011, 46012, 46013, 46014, 46015, 46022, 46025, 46026, 46027, 46028, 46029, 46041, 46042, 46047, 46050, 46053, 46054, 46060, 46061, 46069, 46072, 46075, 46076, 46078, 46081, 46082, 46086, 46088, 46089, 51000, 51003, 51101, 46221, 46214, 46211, 46224, 46215, 46222, 46213, 46239, 46240, 46243, 46232, 44100, and 42099.

REFERENCES

[1] A. Chabchoub, N. Hoffmann, H. Branger, C. Kharif, and N. Akhmediev, "Experiments on wind-perturbed rogue wave hydrodynamics using the peregrine breather model," *Phys. Fluids*, vol. 25, no. 10, 2013, Art. no. 101704.

[2] D. R. Solli, C. Ropers, P. Koonath, and B. Jalali, "Optical rogue waves," *Nature*, vol. 450, no. 7172, pp. 1054–1057, Dec. 2007.

[3] J. H. V. Nguyen, D. Luo, and R. G. Hulet, "Formation of matter-wave soliton trains by modulational instability," *Science*, vol. 356, no. 6336, pp. 422–426, Apr. 2017.

[4] W.-R. Sun, B. Tian, Y. Jiang, and H.-L. Zhen, "Rogue matter waves in a Bose-Einstein condensate with the external potential," *Eur. Phys. J. D*, vol. 68, no. 10, pp. 1–7, Oct. 2014.

[5] K. Manikandan, N. V. Priya, M. Senthilvelan, and R. Sankaranarayanan, "Higher-order matter rogue waves and their deformations in two-component Bose-Einstein condensates," in *Waves in Random Complex Media*, vol. 32, no. 2, pp. 1–20, Aug. 2020.

[6] Z.-Y. Yan, "Financial rogue waves," *Commun. Theor. Phys.*, vol. 54, no. 5, pp. 947–949, Nov. 2010.

[7] D. Li and Y. Chen, *Global Classical Solutions for Nonlinear Evolution Equations*. Harlow, U.K.: Longman Scientific & Technical, 1992.

[8] H.-Q. Zhang and F. Chen, "Rogue waves for the fourth-order nonlinear Schrödinger equation on the periodic background," *Chaos: Interdiscipl. J. Nonlinear Sci.*, vol. 31, no. 2, Feb. 2021, Art. no. 023129.

[9] M. Onorato, D. Proment, G. Clauss, and M. Klein, "Rogue waves: From nonlinear Schrödinger breather solutions to sea-keeping test," *PLOS ONE*, vol. 8, no. 2, Feb. 2013, Art. no. e54629.

[10] A. Ankiewicz, M. Bokaeeyan, and N. Akhmediev, "Infinitely extended complex kdv equation and its solutions: Solitons and rogue waves," *Physica Scripta*, vol. 95, no. 3, Mar. 2020, Art. no. 035201.

[11] C. Qian, J.-G. Rao, Y.-B. Liu, and J.-S. He, "Rogue waves in the three-dimensional Kadomtsev–Petviashvili equation," Chinese *Phys. Lett.*, vol. 33, no. 11, Nov. 2016, Art. no. 110201.

[12] J. Rao, L. Wang, W. Liu, and J. He, "Rogue-wave solutions of the Zakharov equation," *Theor. Math. Phys.*, vol. 193, no. 3, pp. 1783–1800, Dec. 2017.

[13] C. Q. Dai, C. L. Zheng, and H. P. Zhu, "Controllable rogue waves in the nonautonomous nonlinear system with a linear potential," *Eur. Phys. J. D*, vol. 66, no. 4, pp. 1–8, Apr. 2012.

[14] J. M. Dudley, G. Genty, A. Mussot, A. Chabchoub, and F. Dias, "Rogue waves and analogies in optics and oceanography," *Nature Rev. Phys.*, vol. 1, no. 11, pp. 675–689, Nov. 2019.

[15] P. Pokhrel, E. Ioup, M. T. Hoque, M. Abdelguerfi, and J. Simeonov, "Forecasting rogue waves in oceanic waters," in *Proc. IEEE Int. Conf. Mach. Learn. Appl.*, 2020, pp. 634–640.

[16] H. v. Haren, "Grand challenges in physical oceanography," *Frontiers Mar. Sci.*, vol. 5, no. 404, 2018.

[17] S. Mafi and G. Amirinia, "Forecasting hurricane wave height in the gulf of Mexico using soft computing methods," *Ocean Eng.*, vol. 146, no. 1, pp. 352–362, 2017.

[18] J. Mahjoobi and A. Etemad-Shahidi, "An alternative approach for the prediction of significant wave heights based on classification and regression trees," *Appl. Ocean Res.*, vol. 30, no. 3, pp. 172–177, 2008.

[19] J. Mahjoobi, A. Etemad-Shahidi, and M. Kazeminezhad, "Hindcasting of wave parameters using different soft computing methods," *Appl. Ocean Res.*, vol. 30, no. 1, pp. 28–36, 2008.

[20] W. Duan, Y. Han, L. Huang, B. Zhao, and M. Wang, "A hybrid EMD-SVR model for the short-term prediction of significant wave height," *Ocean Eng.*, vol. 124, pp. 54–73, 2016.

[21] B. Kamranzad, A. Etemad-Shahidi, and M. Kazeminezhad, "Wave height forecasting in Dayyer, the Persian gulf," *Ocean Eng.*, vol. 38, no. 1, pp. 248–255, 2011.

[22] M. Ozger, "Significant wave height forecasting using wavelet fuzzy logic approach," *Ocean Eng.*, vol. 37, no. 16, pp. 1443–1451, 2010.

[23] M. Özger and Z. Şen, "Prediction of wave parameters by using fuzzy logic approach," Jul. 2006. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0029801806001065?via%3Dihub

[24] T. Hlophe, H. Wolgamot, A. Kurniawan, P. H. Taylor, J. Orszaghova, and S. Draper, "Fast wave-by-wave prediction of weakly nonlinear unidirectional wave fields," *Appl. Ocean Res.*, vol. 112, 2021, Art. no. 102695. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0141118721001723

[25] J. R. Halliday, D. G. Dorrell, and A. R. Wood, "An application of the fast Fourier transform to the short-term prediction of sea wave behaviour," *Renewable Energy*, vol. 36, no. 6, pp. 1685–1692, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960148110005525

[26] M. Brühl and H. Oumeraci, "Analysis of long-period cosine-wave dispersion in very shallow water using nonlinear Fourier transform based on kdv equation," *Appl. Ocean Res.*, vol. 61, pp. 81–91, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0141118716300669

[27] R. Stuhlmeier and M. Stiassnie, "Deterministic wave forecasting with the Zakharov equation," *J. Fluid Mechan.*, vol. 913, pp. 1–20, 2021.

[28] N. Barton *et al.*, "The navy's earth system prediction capability: A new global coupled atmosphere-ocean-sea ice prediction system designed for daily to subseasonal forecasting," *Earth Space Sci.*, vol. 8, no. 4, pp. 1–28, Sep. 2020.

[29] M. Li and K. Liu, "Probabilistic prediction of significant wave height using dynamic Bayesian network and information flow," *Water*, vol. 12, no. 8, Aug. 2020, Art. no. 2075.

[30] S. Mafi and G. Amirinia, "Forecasting hurricane wave height in gulf of Mexico using soft computing methods," *Ocean Eng.*, vol. 146, pp. 352–362, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0029801817305863

[31] M. Pirhooshyaran and L. V. Snyder, "Forecasting, hindcasting and feature selection of ocean waves via recurrent and sequence-to-sequence networks," Apr. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0029801820304492

[32] S. Londhe, S. Shah, P. Dixit, T. B. Nair, P. Sirisha, and R. Jain, "A coupled numerical and artificial neural network model for improving location specific wave forecast," Aug. 2016. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0141118716300876

[33] Z. Wei, "Forecasting wind waves in the us atlantic coast using an artificial neural network model: Towards an ai-based storm forecast system," Aug. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0029801821010234

[34] H. Hu, A. J. v. d. Westhuysen, P. Chu, and A. Fujisaki-Manome, "Predicting lake erie wave heights and periods using XGBoost and LSTM," Jun. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1463500321000846

[35] R. Lou, W. Wang, X. Li, Y. Zheng, and Z. Lv, "Prediction of ocean wave height suitable for ship autopilot," *IEEE Trans. Intell. Transp. Syst.*, early access, Mar. 2021, doi: 10.1109/TITS.2021.3067040.

[36] P. Bento, J. Pombo, R. Mendes, M. Calado, and S. Mariano, "Ocean wave energy forecasting using optimised deep learning neural networks," Dec. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0029801820312798

[37] A. N. Deshmukh, K. Sandhya, T. B. Nair, P. K. Bhaskaran, and M. Deo, "Neural-network-based data assimilation to improve numerical ocean wave forecast," Apr. 2016. [Online]. Available: https://ieeexplore.ieee.org/document/7460181

[38] X. Zhang, S. Gao, T. Wang, Y. Li, and P. Ren, "Correcting predictions from simulating wave nearshore model via Gaussian process regression," in *Proc. Global Oceans Conf.: Singapore - U. S. Gulf Coast*, 2020, pp. 1–4.

[39] J. Mooneyham, S. C. Crosby, N. Kumar, and B. Hutchinson, "SWRL net: A spectral, residual deep learning model for improving short-term wave forecasts," Dec. 2020. [Online]. Available: https://journals.ametsoc.org/view/journals/wefo/35/6/WAF-D-19-0254.1.xml

[40] N. Rascle and F. Ardhuin, "A global wave parameter database for geophysical applications. part 2: Model validation with improved source term parameterization," *Ocean Model.*, vol. 70, pp. 174–188, Oct. 2013.

[41] A. Vaswani *et al.*, "Attention is all you need," Dec. 2017, *arXiv: 1706.03762*. [Online]. Available: http://arxiv.org/abs/1706.03762

[42] X. Jia *et al.*, "Physics guided RNNs for modeling dynamical systems: A case study in simulating lake temperature profiles," in *Proc. 2019 SIAM Int. Conf. Data Mining*, 2019, pp. 558–566.

[43] P. A. Watson, "Applying machine learning to improve simulations of a chaotic dynamical system using empirical error correction," *J. Adv. Model. Earth Syst.*, vol. 11, no. 5, pp. 1402–1417, 2019.

[44] National Oceanic US Department of Commerce and Atmospheric Administration, "National data buoy center," [Online]. Available: https://www.ndbc.noaa.gov/

[45] "Ifremer wave hindcasts." [Online]. Available: https://forge.ifremer.fr/plugins/mediawiki/wiki/ww3/index.php/En:ifremer_wave_hindcasts

[46] B. Li *et al.*, "Ode transformer: An ordinary differential equation-inspired model for neural machine translation," Apr. 2021, *arXiv: 2104.02308*. [Online]. Available: http://arxiv.org/abs/2104.02308

[47] N. Yadav, A. Yadav, and M. Kumar, "An introduction to neural network methods for differential equations - springer," Mar. 2015. [Online]. Available: https://link.springer.com/book/10.1007/978-94-017-9816-7

[48] A. Kratsios, "The universal approximation property," *Ann. Math. Artif. Intell.*, vol. 89, no. 5–6, pp. 435–469, Jun. 2021.

[49] T. Chen and H. Chen, "Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems," *IEEE Trans. Neural Netw.*, vol. 6, no. 4, pp. 911–917, Jul. 1995.

[50] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Netw.*, vol. 2, no. 5, pp. 359–366, Jan. 1989.

[51] C. Yun, S. Bhojanapalli, A. Rawat, S. J. Reddi, and S. Kumar, "Are transformers universal approximators of sequence-to-sequence functions?," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–23.

[52] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[53] K. M. Kam, "Stationary and non-stationary time series prediction using state space model and pattern-based approach," M.S. thesis, Univ. Texas Arlington, Arlington, TX, USA, 2014.

[54] J. V. Greunen, A. Heymans, C. V. Heerden, and G. v. Vuuren, "The prominence of stationarity in time series forecasting," *J. Stud. Econ. Econ.*, vol. 38, no. 1, pp. 1–16, 2014.

[55] D. A. Dickey and W. A. Fuller, "Distribution of the estimators for autoregressive time series with a unit root," *J. Amer. Stat. Assoc.*, vol. 74, no. 366a, pp. 427–431, 1979. [Online]. Available: https://doi.org/10.1080/01621459.1979.10482531

[56] D. Kwiatkowski, P. C. Phillips, P. Schmidt, and Y. Shin, "Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?," *J. Econ.*, vol. 54, no. 1, pp. 159–178, 1992. [Online]. Available: https://www.sciencedirect.com/science/article/pii/030440769290104Y

[57] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," Feb. 2021, *arXiv: 2101.01169*. [Online]. Available: http://arxiv.org/abs/2101.01169

[58] J.-R. Bidlot, "Intercomparison of operational wave forecasting systems against buoys: Data From ECMWF, MetOffice, FNMOC, MSC, NCEP, MeteoFrance, DWD, BoM, SHOM, JMA, KMA, Puerto Del Estado, DMI, CNR-AM, METNO, SHN-SM January 2015 to December 2015," Eur. Centre Medium-Range Weather Forecasts, Apr. 2016.

[59] J. Bidlot, P. A. E. M. Janssen, and S. Abdalla, "A revised formulation of ocean wave dissipation and its model impact," *ECMWF Tech. Memoranda*, vol. 509, pp. 1–27, 2007.

[60] P. A. Janssen, "Progress in ocean wave forecasting," *ECMWF Tech. Memoranda*, vol. 529, pp. 3572–3594, 2007.

[61] B. Golding, "A wave prediction system for real-time sea state forecasting," *Quart. J. Roy. Meteorological Soc.*, vol. 109, no. 460, pp. 393–416, 1983.

[62] K. M. Wingeart, T. H. C. Herbers, W. C. O'Reilly, P. A. Wittmann, R. E. Jensen, and H. L. Tolman, "Validation of Operational Global Wave Prediction Models With Spectral Buoy Data," in *Proc. 4th Int. Symp. Ocean Wave Meas. Anal.*, Apr. 2012, pp. 590–599,.

[63] J. D. Dykes and W. E. Rogers, "WAVEWATCH III (registered trademark): Transition to naval operations," Jan. 2013. [Online]. Available: https://apps.dtic.mil/sti/citations/ADA592312

[64] J. D. Dykes and W. E. Rogers, "WAVEWATCH III: Transition to naval operations," Nov. 2011. [Online]. Available: https://apps.dtic.mil/sti/citations/ADA555919

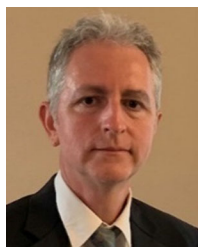
[65] S. Desjardins, J. Mailhot, and R. Lalbeharry, "Examination of the impact of a coupled atmosphere and ocean wave system. Part I: Atmospheric aspects," *J. Phys. Oceanogr.*, vol. 30, no. 2, pp. 402–415, 2000.
[66] R. Lalbeharry, "Evaluation of the CMC regional wave forecasting system against buoy data," *Atmos.-Ocean*, vol. 40, no. 1, pp. 1–20, 2010.
[67] H. L. Tolman *et al.*, "Development and implementation of wind-generated ocean surface wave modelsat NCEP," Apr. 2002. [Online]. Available: https://journals.ametsoc.org/view/journals/wefo/17/2/1520-0434_2002_017_0311_daiowg_2_0_co_2.xml
[68] H. L. Tolman, "Alleviating the garden sprinkler effect in wind wave models," *Ocean Model.*, vol. 4, no. 3–4, pp. 269–289, 2002.
[69] B. Frandon, D. Hauser, and J.-M. Lefevre, "Comparison study of a second-generation and of a third-generation wave prediction model in the context of the semaphore experiment," *J. Atmospheric Ocean. Technol.*, vol. 17, no. 2, pp. 197–214, 2000.
[70] J.-M. Lefevre and P. Cotton, "Chapter 7 ocean surface waves," *Int. Geophys.*, vol. 69, pp. 305–328, 2001.
[71] A. Behrens and D. Schrader, "The wave forecast system of the Deutscher Wetterdienst and the bundesamt fur seeschiffahrt und hydrographique: A verification using ERS-1 altimeter and scatterometer data," *Ocean Dyn.*, vol. 46, no. 2, pp. 131–149, 1994.
[72] L. C. Bender, "Modification of the physics and numerics in a third-generation ocean wave model," *J. Atmospheric Ocean. Technol.*, vol. 13, no. 3, pp. 726–750, 1996.
[73] E. W. Schulz, J. D. Kepert, and D. J. Greenslade, "An assessment of marine surface winds from the australian bureau of meteorology numerical weather prediction systems," *Weather Forecasting*, vol. 22, no. 3, pp. 613–636, 2007.
[74] D. J. Greenslade, E. W. Schulz, J. D. Kepert, and G. R. Warren, "The impact of the assimilation of scatterometer winds on surface wind and wave forecasts," *J. Atmospheric Ocean Sci.*, vol. 10, no. 3, pp. 261–287, 2007.
[75] F. Ardhuin, T. Herbers, K. P. Watts, G. P. van Vledder, R. Jensen, and H. C. Graber, "Swell and slanting-fetch effects on wind wave growth," *J. Phys. Oceanogr.*, vol. 37, no. 4, pp. 908–931, 2007.
[76] K. Ueno and K. Nadao, "The development of the third-generation wave model MRI-III," in *Proc. 8th Int. Workshop Wave Hindcasting Forecasting*, 2004, pp. 1–7.
[77] S. Park, D.-U. Lee, and J.-W. Seo, "Operational wind wave prediction system at KMA," in *Proc. JCOMM Sci. Tech. Symp. Storm Surges*, 2008, pp. 133–150.
[78] K. E. Trenberth, J. T. Fasullo, and J. Mackaro, "Atmospheric moisture transports from ocean to land and global energy flows in reanalyses," Sep. 2011. [Online]. Available: https://journals.ametsoc.org/view/journals/clim/24/18/2011jcli4171.1.xml
[79] D. P. Dee, M. Balmaseda, G. Balsamo, R. Engelen, A. J. Simmons, and J.-N. Thépaut, "Toward a consistent reanalysis of the climate system," Aug. 2014. [Online]. Available: https://journals.ametsoc.org/view/journals/bams/95/8/bams-d-13-00043.1.xml
[80] R. W. Reynolds, N. A. Rayner, T. M. Smith, D. C. Stokes, and W. Wang, "An improved in situ and satellite SST analysis for climate," *J. Climate*, vol. 15, no. 13, pp. 1609–1625, Jul. 2002.
[81] W. Lin, K. Hasenstab, G. Moura Cunha, and A. Schwartzman, "Comparison of handcrafted features and convolutional neural networks for liver MR image adequacy assessment," Nov. 2020. [Online]. Available: https://www.nature.com/articles/s41598-020-77264-y
[82] F. T. Krogh, "Algorithms for changing the step size," Oct. 1973. [Online]. Available: https://www.proquest.com/openview/406d0b16d544e48fc7a70e06efa1c5f3/1?pq-origsite=gscholar&cbl=666303
[83] F. M. McNeill and E. Thro, *Fuzzy Logic: A Practical Approach*. Cambridge. MA, USA: Academic Press Professional, Inc., 1994.
[84] B. Bercu, "Weighted estimation and tracking for armax models," *SIAM J. Control Optim.*, vol. 33, no. 1, pp. 89–106, Jan. 1995.
[85] G. Reikard and W. E. Rogers, "Forecasting ocean waves: Comparing a physics-based model with statistical models," *Coastal Eng.*, vol. 58, no. 5, pp. 409–416, May 2011.
[86] J. Rychtář and D. T. Taylor, "Estimating the sample variance from the sample size and range," *Statist. Med.*, vol. 39, no. 30, pp. 4667–4686, Sep. 2020, doi: 10.1002/sim.8747.
[87] S. Singh and A. H. Joarder, "Estimation of finite population variance using random non-response in survey sampling," *Metrika*, vol. 47, no. 1, pp. 241–249, Jan. 1998, doi: 10.1007/bf02742876.
[88] M. Loeve, *Probability Theory I* (Graduate Texts in Mathematics), 4th ed. Berlin, Germany: Springer-Verlag, 1977. [Online]. Available: https://www.springer.com/gp/book/9780387902104
[89] M. Held, "Deriving skewness and excess kurtosis of the sum of IID random variables," Feb. 2011, no. ID 1758662. [Online]. Available: https://papers.ssrn.com/abstract=1758662
[90] D. Hauser *et al.*, "COST action 714, measuring and analysing the directional spectra of ocean waves. office for official publications of the European communities," 2005. [Online]. Available: https://repository.oceanbestpractices.org/handle/11329/1303
[91] N. Mori, M. Onorato, and P. A. Janssen, "On the estimation of the kurtosis in directional sea states for freak wave forecasting," *J. Phys. Oceanogr.*, vol. 41, no. 8, pp. 1484–1497, Mar. 2011, doi: 10.1175/2011jpo4542.1.
[92] A. J. Kuik, G. P. v. Vledder, and L. H. Holthuijsen, "A method for the routine analysis of pitch-and-roll buoy wave data," *J. Phys. Oceanogr.*, vol. 18, no. 7, pp. 1020–1034, Jul. 1988.

Pujan Pokhrel is currently working toward the Ph.D. degree in computer science with the Canizaro Livingston Gulf States Center for Environmental Informatics, University of New Orleans, New Orleans, LA, USA.

His research interests include inverse problems, machine learning, artificial intelligence, and fluid dynamics.

Elias Ioup received the Ph.D. degree in engineering and applied science from The University of New Orleans, New Orleans, LA, USA, in 2007. He is currently a Computer Scientist and the Head of the Geospatial Computing Section, U.S. Naval Research Laboratory, Washington, DC, USA.

His research interests include high-performance geospatial data processing, geospatial and environmental web services, and geospatial data visualization.

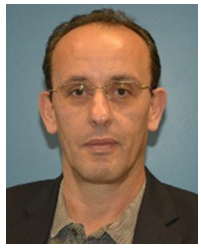


Julian Simeonov received the M.S. degree in meteorology from Sofia University, Sofia, Bulgaria, in 1996, and the Ph.D. degree in oceanography from Florida State University, Tallahassee, FL, USA, in 2002.

He was a Postdoctoral Associate with Florida State University from 2003 to 2007 and an American Society for Engineering Education Postdoctoral Associate with the U.S. Naval Research Laboratory (NRL), Stennis Space Center, Hancock County, MS, USA, from 2008 to 2010. Since 2011, he has been an Oceanographer with the Sediment Dynamics Section, NRL Seafloor Sciences Branch and the Principal Investigator for five basic and applied research projects. He is the author of 24 peer-reviewed publications and one US patent. His research interests include computational and theoretical fluid dynamics, multiphase modeling of sediment transport, and statistical inference and inverse methods in geophysics.

Md Tamjidul Hoque received the Ph.D. degree in information technology from Monash University, Melbourne, VIC, Australia, in 2008.

He is currently an Associate Professor with the Computer Science Department, University of New Orleans, New Orleans, LA, USA. From 2011 to 2012, he was a Postdoctoral Fellow with Indiana University-Purdue University Indianapolis, Indianapolis, IN, USA. From 2007 to 2011, he was a Research Fellow with Griffith University, Brisbane, QLD, Australia. His current research interests include deep/machine learning, evolutionary computation, and artificial intelligence, applying toward hard optimization problems, especially for bioinformatics problems such as protein structure-prediction, disorder predictor, and energy function.



Mahdi Abdelguerfi received the Ph.D. degree from Wayne State University, Detroit, MI, USA, in 1987.

From 1987 to 1989, he was an Assistant Professor with the University of Detroit, Detroit, MI, USA. From 1989 to 1995, he was an Associate Professor with the Department of Computer Science, The University of New Orleans, New Orleans, LA, USA, where he was the Department Chair from 1998. He is currently the Chairman of the Computer Science Department, The University of New Orleans. He is the Founder and Executive Director of the Canizaro Livingston Gulf States Center for Environmental Informatics (GulfSCEI), UNO.