IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 6, 2021, accepted July 2, 2021, date of publication July 7, 2021, date of current version July 16, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3095339

Advanced Machine Learning Techniques for Predicting Nha Trang Shorelines

CHENG YIN[1], LE THANH BINH[2], DUONG TRAN ANH[3], SON T. MAI[1], ANH LE[4], VAN-HAU NGUYEN[5], VAN-CHIEN NGUYEN[6], NGUYEN XUAN TINH[7], HITOSHI TANAKA[7], NGUYEN TRUNG VIET[8], LONG D. NGUYEN[9], (Member, IEEE), AND TRUNG Q. DUONG[1], (Senior Member, IEEE)

[1]Queen's University Belfast, Belfast BT7 1NN, U.K.
[2]Vietnam Hydraulic Engineering Consultants Corporation—JSC (HEC), Hanoi, Vietnam
[3]Ho Chi Minh City University of Technology (HUTECH), Ho Chi Minh City, Vietnam
[4]University of Transport, Ho Chi Minh City 700000, Vietnam
[5]Hung Yen University of Technology and Education, Hải Dương 17000, Vietnam
[6]Hanoi University of Science and Technology, Hanoi 10000, Vietnam
[7]Tohoku University, Sendai 980-8579, Japan
[8]Thuyloi University, Hanoi 10000, Vietnam
[9]Duy Tan University, Da Nang 810000, Vietnam

Corresponding authors: Trung Q. Duong (trung.q.duong@qub.ac.uk) and Nguyen Trung Viet (nguyentrungviet@tlu.edu.vn)

This work was supported by a Researcher Links grant, ID 527612186, under the Newton Fund partnership. The grant is funded by the UK Department for Business, Energy and Industrial Strategy and delivered by the British Council, for further information, please visit www.newtonfund.ac.uk.

ABSTRACT Nha Trang Coast is located in the South Central Vietnam and the coastal erosion has occurred rapidly in recent years. Hence it is crucial to accurately monitor the shoreline changes for better coastal management and reduction of risks for communities. In this paper, we explored a statistical forecasting model, Seasonal Auto-regressive Integrated Moving Average (SARIMA), and two Machine Learning (ML) models, Neural Network Auto-Regression (NNAR) and Long Short-Term Memory (LSTM), to predict the shoreline variations from surveillance camera images. Compared to the Empirical Orthogonal Function (EOF), the most common method used for predicting shoreline changes from cameras, we demonstrate that the SARIMA, NNAR and LSTM models outperform the EOF model significantly in terms of prediction accuracy. The forecasting performance of the SARIMA model, NNAR model and LSTM model is comparable in both long and short-term predictions. The results suggest that these models are highly effective in detecting shoreline changes from video cameras under extreme weather conditions.

INDEX TERMS Nha Trang coast, shoreline prediction, statistical forecasting model, machine learning, EOF, SARIMA, NNAR, LSTM.

I. INTRODUCTION

Coastal erosion is a natural phenomenon driven by numerous complex physical processes of wave action, tidal regime and sediment transportation. Over 70 % of the worldwide shorelines experienced erosion [1]. Understanding the physical processes and the intervention of human activities is crucial for efficient coastal management. Recently, Nha Trang Coastline suffers from accelerated erosion due to strong northeast monsoon with forceful winds and high waves. The landscape of the coast is damaged due to climate change and sea level rise, and the local economy is threatened by the decrease of the number of tourists and the high amount of reconstruction costs after coastal disasters. Several past studies worked on the erosion evolution of Nha Trang Coast, e.g. [2]–[4]. Though these studies documented the insights of seasonal changes and sediment transports of Nha Trang Coast, the lack of effective prediction models remains a challenge for coastal management. Therefore, decision makers desire a good forecasting model to predict shoreline positions so that they can make master plans in advance to mitigate disasters and develop coastal economy in the local area.

Coastal design methods are classified into three categories, physical models, numerical models and field data analysis [5]. Physical models analyze and predict shoreline changes through reproducing the miniature of a physical

The associate editor coordinating the review of this manuscript and approving it for publication was Xianzhi Wang.

system. However, transferring a realistic coastal profile into a laboratory model is challenging due to the scale effects. More importantly, data collection and model validation are laborious and time-consuming, so this method may not valid for long-term prediction [6]. A numerical model represents a physical system via mathematical functions. For instance, a numerical model, GENEeralized SImulating Shoreline change model (GENESIS), has been developed to simulate and forecast long-term shoreline variation [7]. The significant wave height and significant wave period were the inputs of a GENESIS model, and the simulation outcomes were used to forecast future shoreline positions. Meanwhile, Nguyen *et al.* [3] analyzed shoreline changes through a numerical one-line theory model incorporating the wave conditions. Besides, Thanh *et al.* [4] conducted Empirical Orthogonal Function (EOF) to explore the dominant variation process of Nha Trang Coast. Field data analysis has been examined as a reliable methodology for shoreline forecasting [6], [8]. Different from numerical models, field data analysis is based on data science and constructs models according to data characteristics rather than physical theorems. Linear regression methods have been implemented to analyze the shoreline positions with the GPS data collected from the field [9]. Non-linear auto-regression neural networks have also been applied to predict the shoreline changes with measurement of shoreline profiles [6]. These models exhibited remarkable predictive ability, although there were several underestimates or overestimates during the prediction periods.

According to the above literature, both numerical models and field data analysis are desirable options to carry out shoreline predictions. However, the methodology with superior predictive ability is still ambiguous for academia. It is essential to assess the performance of these forecasting models and identify the one with the best predictive ability. In this study, we achieve empirical results based on the data collected from Nha Trang Coast. Our work provides guidance for academics seeking appreciative forecasting models in relevant research.

Conventional EOF model, a widely used numerical model, is developed as the benchmark prediction model in this study. EOF model was firstly introduced by Lorenz in 1956 [10] and has been extensively applied to explore the spatial and temporal changes in oceanography, meteorology and climate science fields. For instance, EOF method has been applied to analyze summer precipitation variability [11], predict subseasonal climate changes [12], [13] and estimate wave height in extreme weather [14]. In our study area, Nha Trang Coast, Thanh *et al.* [4] analyzed shoreline variations via EOF model. In line with this research, we apply this conventional EOF model to extract dominant temporal components and then predict shoreline variations.

Next, we examine the predictive abilities of three field data analysis models, including a statistical forecasting model and two non-linear Machine Learning (ML) models. The statistical model termed as Seasonal Auto-Regressive Integrated Moving Average (SARIMA) model is an extension of Auto-Regressive Integrated Moving Average (ARIMA) model [15], but analyzes the time series with additional seasonal components [16]. SARIMA model is an effective tool for seasonal time series forecasting. For instance, SARIMA model was used to forecast sea level [17], wind speed [18] and wave height [19]. Those studies all supported that SARIMA model yields satisfactory forecasting performance.

Apart from the statistical forecasting model, we further consider Neural Network Auto-Regression (NNAR) which has the ability to execute complicated non-linear functions. Lagged data is used to train the NNAR model and then the model generates prediction values based on the trained networks. The major difference between NNAR model and ARIMA model is that NNAR model does not restrict the parameters to secure the stationary time series [20]. This model has been extensively used by tourism studies [21], [22]. In addition, a comparison of forecasting performance between ARIMA and NNAR was conducted in [20]. Their results validated that NNAR outperforms ARIMA for all observed time series in the prediction of Water Treatment Plant (WTP) influent characteristics.

Lastly, we apply another non-linear ML technique, Long Short-Term Memory (LSTM) model, which has efficient ability to remember the information of long periods [23]. LSTM model has been used to analyze wind speed in the coastal belts of India peninsula suffering from extreme weather such as storms [24]; forecast sea level around the Korean Coast, achieving good performance with R over 0.85 [25]; analyze wave height, yielding outperformance with a low MAPE at 5.15 % for 1-h prediction, which significantly outperforms Simulating WAves Nearshore (SWAN) model [26]. Research of [27] compared the performance of SARIMA, NNAR and LSTM model for early detection of disease incidents and found that LSTM model is superior to SARIMA and NNAR. Similar results were also concluded in [28] that LSTM model outperforms ARIMA model. The remarkable predictive ability of LSTM indicated by previous research motivates us to involve this model as one of the candidates to predict shoreline changes.

Although previous studies proposed that sophisticated ML models with powerful computational abilities outperform statistical methods [20], [27], [28], the superiority of ML models vanishes in some cases. For instance, simple ARIMA model outperformed NNAR model when predicting monthly tourist arrivals [21]. A more comprehensive comparison was conducted in [29]. After assessing the forecasting performance of eight statistical methods and ten ML methods in a sample of 1045 monthly time series, the authors concluded that statistical methods perform better than ML methods in both single-step and multi-step predictions. These outcomes suggest that sophisticated ML models fail to persistently be superior to simple time-series forecasting models. These findings motivate us to implement not only sophisticated ML techniques but also simple methods like SARIMA model for shoreline prediction.

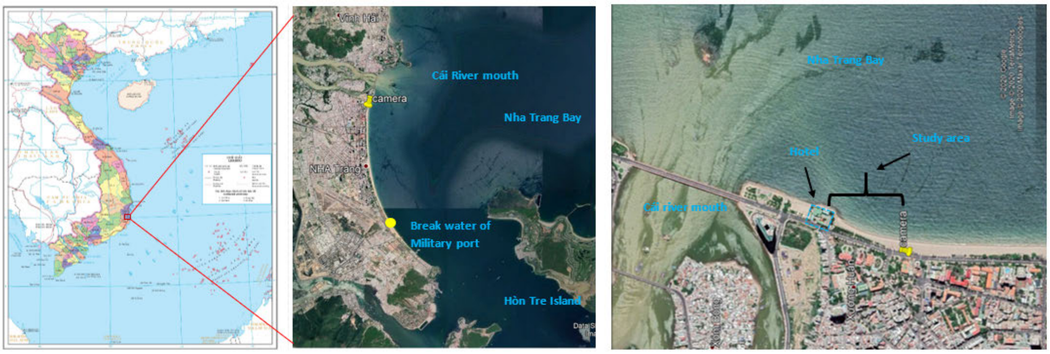


FIGURE 1. Nha Trang coast extends from Cai river mouth in the north to the breakwater of the military port in the south; The camera system was installed between the river mouth and the breakwater. The study area focuses on the north of the camera system until a beach-side hotel, at a length of around 300m.

Contributions: This paper aims to explore effective models for predicting shoreline changes of Nha Trang Coast via surveillance camera images including both statistical and advanced ML models including NNAR and LSTM. The main contributions are summarized as below:

- Current camera surveillance system to monitor coastal changes usually suffers from extreme weather conditions, which leads to plenty of noisy and missing data due to a variety of factors, e.g. signal transmission disruptions [4]. These missing data causes a severe impact on methods for predicting shoreline changes. Hence, we explore four extensively used imputation methods to impute missing values before feeding data to prediction models, namely zero imputation, seasonal adjustment methods including linear interpolation and Last Observation Carried Forward (LOCF), and K-Nearest Neighbors (KNN) algorithm. We show that seasonal adjusted linear interpolation is the best method for imputing missing coastal data.
- While existing studies for Nha Trang only rely on EOF for predicting the changes due to its robustness to missing data, e.g. [4], we aim at exploring more effective techniques including 1) a statistical forecasting model, SARIMA model, and 2) ML models, NNAR model and LSTM model to analyze shoreline variation of Nha Trang Coast. We show that these methods acquire significant performance boosts over EOF when being combined with suitable data imputation techniques.
- We further demonstrate that the performance of SARIMA, NNAR and LSTM in long-term coastal change predictions (up to 50 days ahead). The results suggest that all these methods can be effectively used for monitoring coastal changes in long-term.

The remainder of the paper is organized as follow. Study areas and data are presented in Section II. Section III reports the forecasting methods. Then, results are summarized in Section IV. Lastly, we conclude our findings in Section V.

II. STUDY AREAS AND DATA

A. STUDY AREA

Nha Trang Coast, located at Nha Trang Bay, is one of the most famous coasts around the world. The length of the coast is around 4.3 km, starting from Cai River mouth in the north to a breakwater of the military port in the south, as shown in Fig. 1. Nha Trang Coast experiences heavy northeast monsoon. The big Hon Tre island and other small islands in the southeast protect the coast from waves from the south and southeast direction. Recently, coastal erosion has been happening seriously in Nha Trang Bay due to the monsoon and climate changes [30].

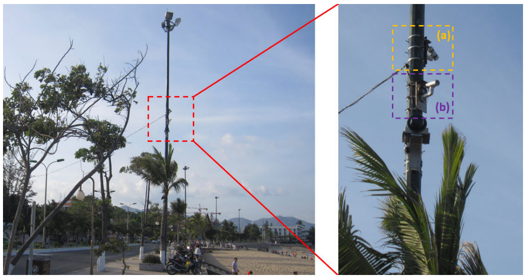


FIGURE 2. Two cameras are mounted on the electric pole, (a) faced the north of the Nha Trang coast and (b) recorded the southern area.

To investigate shoreline variations, it is necessary to monitor the shoreline in long-term first. Different from conventional monitoring techniques such as aerial and satellite that can only capture a single photo, video monitoring techniques attract more interests due to the capability of providing continuous images. Therefore, a video camera system was deployed to analyze the shoreline variation on Nha Trang Coast as described in Fig. 2. The whole system contained two cameras monitoring the entire coast. The northern camera (a) is 10.45m above the mean sea level and the southern camera (b) is 9.67m above the mean sea level. However, the second camera was damaged by a severe thunderbolt in July 2014, so we only apply data from the northern area of the coast to our models.

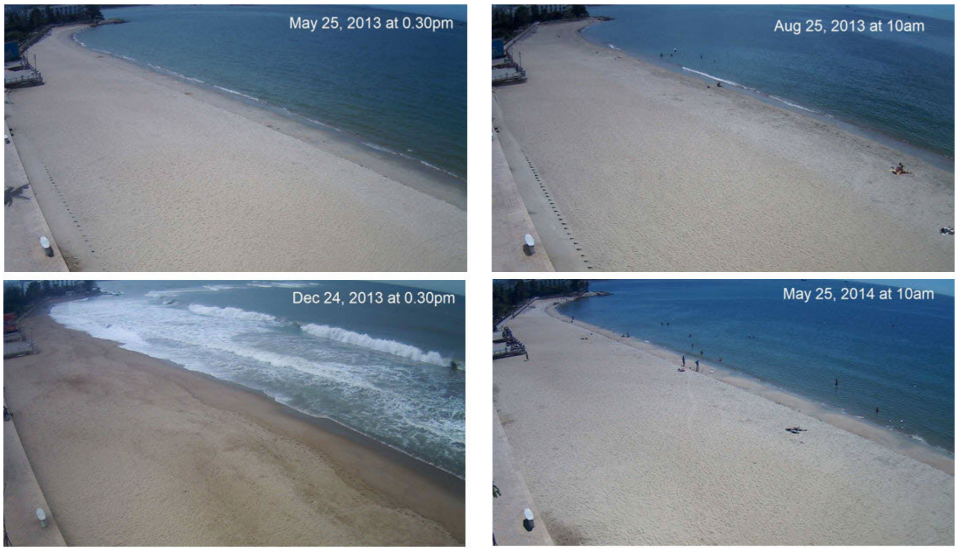


FIGURE 3. Some raw images taken from the camera at different times.

B. DATA COLLECTION AND ANALYSIS

Images captured from the coastal video-camera monitoring system deployed at Nha Trang Coast are considered as raw data in this research. Such images were taken every second from 06:00 am to 05:15 pm every day. Fig. 3 presents raw images captured from the camera and it is difficult to delineate the shoreline of the land-water boundary directly from raw images. Thus, a suitable method for image analysis should be used to extract shorelines from images automatically.

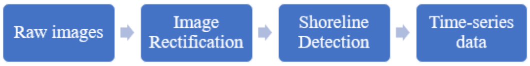


FIGURE 4. Image analysis processing.

The image analysis process in [30] is applied to extract shoreline coordinates from raw image data. As shown in Fig. 4, the process consists of two steps: 1) image rectification with projective transformation and 2) shoreline detection based on changing color intensity between wet and dry sand sides.

For the first step, fifteen-minute averaged images generated from averaged cross-shore positions are transformed to reduce the impact of the alongshore variation. In addition, to minimize the influence of sea level changes, the fifteen-minute averaged images were collected at the time of mean sea level, and 12 Ground Control Points (GCPs) were chosen along the sea dyke with almost the same elevation. Based on those GCPs, the original oblique fifteen-minute averaged images are transformed to horizontal images. Then the fifteen-minute averaged images are corrected from pixel coordinates to real world coordinates system following the research in [31]. In the real world coordinate, the point located at 303959m East and 1355573m North on World Geodetic System (WGS84) is chosen as the original point $(x, y) = (0, 0)$. Positive x-axis denotes the alongshore distance from the original point towards the south, and positive y-axis indicates the cross-shore distance towards the sea as shown in Fig. 5.

In the second step, shoreline coordinates are detected based on the method presented in the research [32]. Basically, an effective shoreline is a boundary between wet and dry sand sides, indicated by changing colors between the two sides. Therefore, we use the gradient maxima of the images to locate the shoreline positions.

Since the step to extract shoreline positions from projected time-averaged images does not consider the elevation of sea level, they need to be corrected based on the geometric relationship among the shoreline position, camera position, sea dyke and sea level. In particular, the shoreline coordinates can be transformed by using the following equations:

$$\bar{x_s} = x_c + (x_s - x_c)\frac{z_s - z_c}{z_d - z_c}, \quad (1)$$

$$\bar{y_s} = y_c + (y_s - y_c)\frac{z_s - z_c}{z_d - z_c}, \quad (2)$$

where $(x_s,\ y_s,\ z_s)$ indicates the detected shoreline position, $(\bar{x_s},\ \bar{y_s},\ \bar{z_s})$ represents the corrected shoreline position, (x_c, y_c, z_c) refers to the camera position, z_d denotes the elevation of sea dyke, z_s is the sea level and $\bar{z} = z_s$.

After image processing, we obtain time dynamic cross-shore distances. Such data, presented in the form of time-series data, will be used as input for the forecasting models in Section III.

Next, we verify the extracted results of the image analysis by using Google Earth images. Fig. 5 shows the comparison of the extracted shoreline from the video camera and Google Earth image on 3 March 2014 and our extracted shoreline matches the captured shoreline in Google Earth image well.

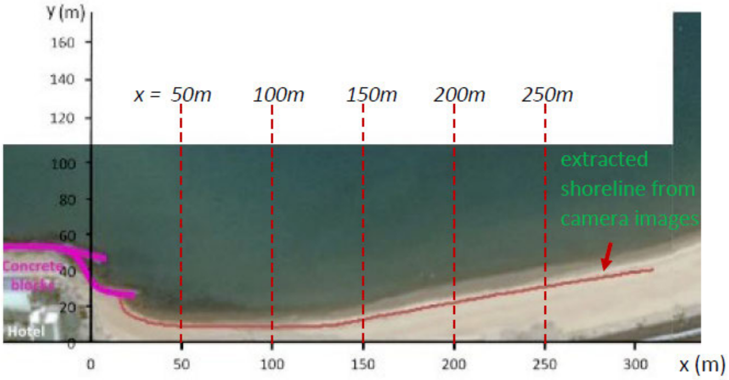


FIGURE 5. Comparison of shoreline position extracted from camera image and Google earth image on 3 March 2014.

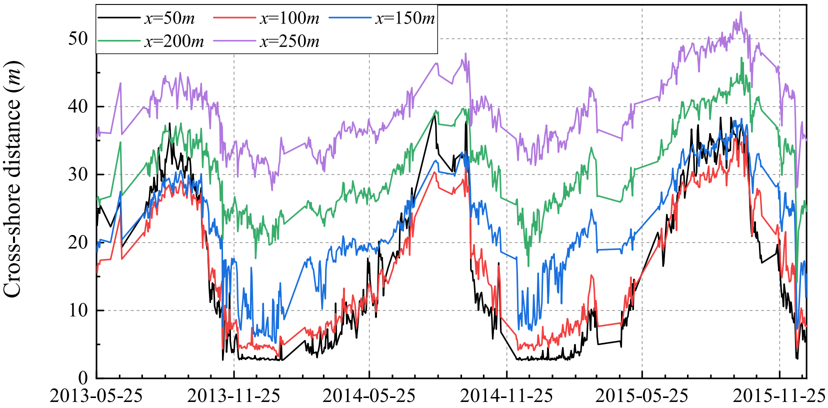


FIGURE 6. Shoreline positions monitored from 25 May 2013 to 31 December 2015 at five positions $x = 50m$, $x = 100m$, $x = 150m$, $x = 200m$ and $x = 250m$.

In this study, we choose five studying points at $x = 50m$, $100m$, $150m$, $200m$ and $250m$ to represent the shoreline situations of the entire coast.

C. DATA EXPLORATION

Daily shoreline positions of five above mentioned locations along the coast from 25 May 2013 to 31 December 2015 are selected as the data used in this research. The variations of cross-shore distance at five studying points are plotted in Fig. 6 and the definition of the coordinate is introduced in Fig. 5. In Fig. 6, cross-shore distances of the five points show similar variation trends but different altitudes.

To further investigate seasonal shoreline variations of Nha Trang Coast, the seasonal analysis at both daily and monthly level are shown in Fig. 7, where position $x = 50m$ is presented as an example since all positions exhibit similar seasonal patterns as shown in Fig. 6. Fig. 7 shows that shoreline advanced from May to September, reaching the peak cross-shore distance in September due to calmer wave during non-monsoon period. Then the shoreline retreated sharply from October to December with the lowest cross-shore distance and remained low level until April. This phenomenon stems from the strong northeast monsoon during winter period from October to April [33].

Fig.8 plots the Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF) to show the seasonality and fluctuations of the time series. According to Fig. 6, the original series is non-stationary, therefore, ACF and PACF graphs are plotted after seasonal differencing with lag 30 due to the daily frequency. ACF shows a significant spike at lag 30 and almost a significant spike at lag 60. Similarly, PACF also show spikes at lag 30 and lag 60. Therefore, the desirable seasonal period of our series is chosen as 30-day.

III. METHODS

In this section, we first introduce the set of training and testing samples. Then, four performance metrics are applied to assess the predictive ability of conducted models. We focus on the prediction of shoreline changes via different methods including SARIMA, NNAR and LSTM. The conventional EOF model is used as the baseline prediction model. In addition, two forecasting strategies, including single-step and multi-step predictions, are applied.

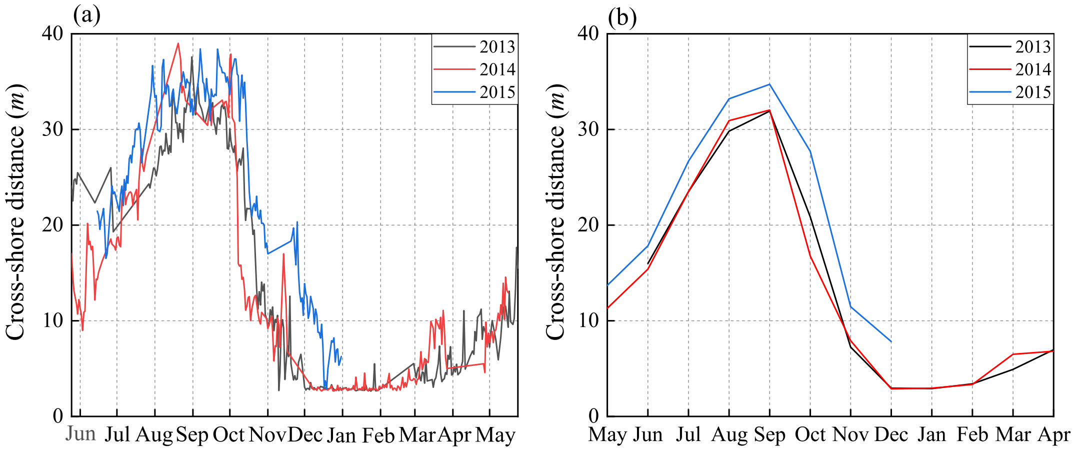


FIGURE 7. Seasonal analysis with (a) daily data and (b) average monthly data at $x = 50m$.

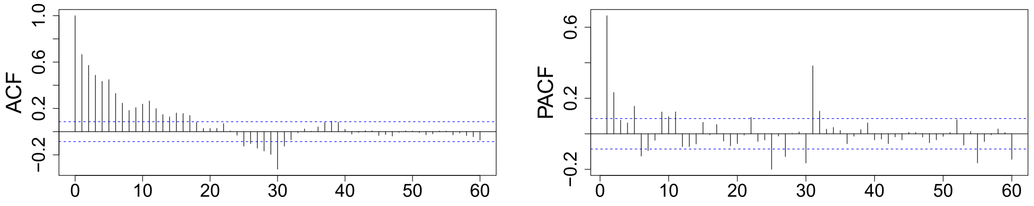


FIGURE 8. ACF and PACF plots with seasonal differenced data at $x = 50m$.

A. DATA PREPROCESSING

The first two-year data (25 May 2013 to 24 May 2015) are applied as training data to serve the models introduced in Section III-D, and the remaining data (from 16 June 2015 to 31 December 2015) is used as testing data to examine the predictive ability of various models.

Some raw images are discarded as the camera fails to capture high quality images during extreme weather, e.g., typhoon and fog. Thus, we have many missing data (around one third of the data). These missing data must be handled before we can implement models like SARIMA. There are two methods to handle missing data: 1) removing the missing values, and 2) filling the missing values based on appropriate imputation approaches. Removing missing records will produce non-continued time series and cause the loss of information. Therefore, it is more desirable to apply an appropriate imputation method to fill the missing observations rather than removing them. Both statistical [34] and ML methods [35] are potential data imputation approaches.

1) ZERO IMPUTATION

A simple and basic filling approach is to impute zero to missing values.

2) SEASONAL TIME SERIES SPECIFIC METHODS

Previous studies agreed that seasonal time series can be decomposed into three terms [36], as:

$$Y_t = T_t + S_t + n_t, \tag{3}$$

where T_t denotes the trend, S_t represents the seasonality, and n_t refers to the noise. The shoreline variations have seasonality as discussed in Section II, so the extraction of the seasonal components and imputation to the de-seasonalized data are carried out with following methods:

- *LOCF:* Replacing missing values with previous non-missing values [37];
- *Linear Interpolation:* An approach of curve fitting via linear polynomials with known values to estimate values of unknown ones [38].

After imputing the non-seasonal term, we then add the seasonal components back [36].

3) ML METHODS

ML algorithms have been widely used for data imputation, and KNN is one of them which has been proven as a robust data imputation method [39]. KNN could estimate a missing point in terms of the values of its closest K neighbours.

B. PERFORMANCE METRICS

To achieve robustness results, four extensively performance metrics (R, MAE, RMSE and MAPE) are used to evaluate prediction accuracy of the models.

R indicates the similarity between actual values and predicted values. For a given model, a higher R represents a better forecasting ability. When R equals one, predicted values are entirely the same as real values. Next, MAE indicates the average differences between predictions and actual values, and RMSE represents the square root of the average squared errors between predictions and actual values. MAE simply averages the absolute errors, therefore, each error contributes to MAE in proportion to the absolute value of each error. Different from MAE, RMSE assigns higher weights to large errors since the differences between actual values and predictions are squared first and then averaged. Therefore, large errors have greater impacts on RMSE than MAE. When RMSE is close to MAE, all errors between predictions and actual values have similar magnitudes. However, when RMSE is dramatically larger than MAE, there are greater variances among errors. RMSE and MAE are used together to evaluate the forecasting performance in terms of large errors in this study. Lastly, MAPE, a statistical metric to measure the forecasting accuracy reported as a percentage, is also used. The formulas of four performance metrics are shown as:

$$\begin{aligned} \mathrm{R} =& \left[n\sum_{i=1}^{n}(A_iB_i) - (\sum_{i=1}^{n}A_i)(\sum_{i=1}^{n}B_i) \right] \\ &\times \frac{1}{\sqrt{n\sum_{i=1}^{n}A_i{}^2 - (\sum_{i=1}^{n}A_i)^2}} \\ &\times \frac{1}{\sqrt{n\sum_{i=1}^{n}(B_i)^2 - (\sum_{i=1}^{n}B_i)^2}}, \end{aligned} \tag{4}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|A_i - B_i|, \tag{5}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(A_i - B_i)^2}, \tag{6}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - B_i}{A_i}\right|, \tag{7}$$

where n is the number of observations, A_i is the real value and B_i is the predicted value. In addition, when assessing the performance of these models, we use the original data without missing data imputation to ensure fair comparisons.

C. FORECASTING STRATEGIES

Two major forecasting procedures, single-step prediction and multi-step prediction, are extensively used to solve time-series prediction problems. The rationales of both forecasting strategies are detailed as below. Assuming a finite segment of previous values up to time t are available as the training sample, shown as $x_t, x_{t-1}, x_{t-2} \ldots, x_1$. Similarly, the original values of the testing sample are $x_{t+1}, x_{t+2}, x_{t+3}, \ldots, x_{t+m}$, where m refers to the size of the testing sample. The single-step prediction only predicts one value out of the known data, and no feedback is employed to pursue the predictions. Differently, multi-step prediction is a recursive procedure that predicted data is used to apply subsequent predictions [40]. For instance, if a model with three inputs is designed to forecast m values in the future, as $y_{t+1}, y_{t+2}, y_{t+3}, \ldots, y_{t+m}$, the procedure of a single-step prediction can be expressed as:

$$\begin{aligned} y_{t+1} &= f[x_t, x_{t-1}, x_{t-2}], \\ y_{t+2} &= f[x_{t+1}, x_t, x_{t-1}], \\ y_{t+3} &= f[x_{t+2}, x_{t+1}, x_t], \\ &\vdots \\ y_{t+m} &= f[x_{t+m-1}, x_{t+m-2}, x_{t+m-3}], \end{aligned} \tag{8}$$

where, $f[\bullet]$ represents the model function determined by the training sample. By contrast, a multi-step prediction re-uses the predictions as additional data to forecast future values as:

$$\begin{aligned} y_{t+1} &= f[x_t, x_{t-1}, x_{t-2}], \\ y_{t+2} &= f[y_{t+1}, x_t, x_{t-1}], \\ y_{t+3} &= f[y_{t+2}, y_{t+1}, x_t], \\ &\vdots \\ y_{t+m} &= f[y_{t+m-1}, y_{t+m-2}, y_{t+m-3}]. \end{aligned} \tag{9}$$

The major distinction between these two procedures is that single-step only uses the original data, whereas multi-step starts with the training sample but overlap the training sample and the predicted values [40]. Therefore, the forecasting error of the multi-step prediction is large as the prediction error at each step is the cumulative error of previous steps.

To avoid large cumulative prediction errors, we implement the rolling-window multi-step prediction to forecast m values in the future. First, the number of prediction steps is defined as K. Second, we implement a multi-step prediction to predict y_{t+K} with the training sample, $x_t, x_{t-1}, \ldots, x_1$, as model inputs. Third, we predict y_{t+K+1} based on the same model, whereas the inputs are updated to $x_{t+1}, x_t, \ldots, x_1$. Then, we update the inputs as $x_{t+2}, x_{t+1}, \ldots, x_1$ to predict y_{t+K+2}. We repeat this procedure until we achieve y_{t+m}. [1]

D. FORECASTING MODELS

This section presents the main features of the four forecasting models. First, a well-known conventional method in meteorology and oceanography fields is studied, namely EOF model, to extract main dominant components of the shoreline and forecast the shoreline variations. Second, a statistical forecasting model SARIMA which has been used in many practically forecasting tasks is investigated. Finally, we consider non-linear model as shoreline changes are non-linear in general [6]. Two ML methods are considered to improve the accuracy of our study on shoreline changes, namely NNAR and LSTM. In this paper, the best model of each forecasting method is the one with the minimum MAPE over the testing period.

1) EOF MODEL

EOF model is an important approach to analyze variations and extract key patterns. In this study, we follow the research

[1] To fix the prediction steps, we exclude the first $(K-1)$ predictions from the forecasting sample.

TABLE 1. Coefficients of fourier series.

Coefficients	Value
a_0	0.47
a_1	-11.09
b_1	-13.19
a_2	6.64
b_2	-4.32

in [4] where EOF model was used to analyze the shoreline variation of Nha Trang Coast and the forecasting performance of EOF model will be used as the benchmark.

In EOF model, shoreline position data is decomposed into two dimensions, i.e., spatial eigenfunction, $e(x)$, and temporal eigenfunction, $c(t)$. The dominant spatial feature is determined by the first spatial eigenfunction $e_1(x)$, which describes the trend of shoreline variation at different coast positions from $x = 50m$ to $x = 250m$. The temporal eigenfunction $c_1(t)$ explains the shoreline variation in different seasons, where t indicates time. The variation of average shoreline position $\bar{y}(x)$ is derived from measured data $y_m(x, t)$, and is given as:

$$y(x, t) = y_m(x, t) - \bar{y}(x). \tag{10}$$

Next, the EOF analysis is displayed as [10] :

$$y(x, t) = \sum_{i=1}^{\infty} e_i(x)c_i(t), \tag{11}$$

where i indicates the $i - th$ component.

The contribution rate Q_i corresponds to the eigenvalues λ_i is given as [10]:

$$Q_i = \frac{\lambda_i}{\sum_{n=1}^{k} \lambda_n}. \tag{12}$$

Based on shoreline data, the contribution rate of the first component, $i = 1$, is 95.20 % in our EOF model, and the total contribution rate of the remaining components is only 4.80 %, much lower than that of the first component. Therefore, the first component responses for nearly the entire of shoreline variation, and other components are less representative and negligible [4]. The first spatial eigenfunction $e_1(x)$ can be achieved by the spectral decomposition. This terms reflects the tendency of shoreline variation at the studied region and is exclusively determined by the position.

Through applying Fourier series, the first temporal eigenfunction, $c_1(t)$, can be decomposed as:

$$c_1(t) = \frac{a_0}{2} + \sum_{i=1}^{2} [a_i \cos(iwt) + b_i \sin(iwt)]. \tag{13}$$

where a_0 is the intercept, and w is the angular frequency ($w = 2\pi/365$ day $^{-1}$). In terms of our data universe, the estimated coefficients are shown in Table 1.

Therefore, a simulation model is developed based on the first spatial and temporal component, and it is utilized to forecast shoreline variation. The equation for shoreline prediction using EOF is presented as:

$$y_m(x, t) = \bar{y}(x) + e_1(x)c_1(t). \tag{14}$$

2) SARIMA MODEL

For time-series prediction, ARIMA model is a remarkable algorithm used by a majority of academic studies. ARIMA is a generalization of Auto-Regressive Moving Average (ARMA) model introduced by Whittle [15]. Mckerchar and Delleur [16] further developed SARIMA model that incorporates seasonal effects of input variables. SARIMA model is normally presented as SARIMA$(p, d, q) \times (P, D, Q)_s$. The p, d, q refer to the non-seasonal polynomial orders of auto-regressive, integrated, and moving average, respectively. The P, D, Q represent the seasonal orders, and s is the seasonal period. Technically, the model is presented as:

$$\phi_p(B)\Phi_P(B^s)\Delta d\Delta D^s y_t = \theta_q(B)\Theta_Q(B^s)\epsilon_t, \tag{15}$$

where y_t is the input variable (the cross-shore distance in this study); $\phi_p(B)$ and $\theta_q(B)$ are the seasonal auto-regression (AR) and moving average (MA) components, respectively; $\Phi_P(B^s)$ and $\Theta_Q(B^s)$ represent the non-seasonal components. The differentiation factors, Δd and ΔD^s, are employed to mitigate the seasonal and non-seasonal non-stationary.

A large number of SARIMA models based on different combinations of non-seasonal and seasonal orders are evaluated. Despite various metrics are used to identify the out-performed models with data preprocessing, (e.g., ACF, PACF, Schwarz information criterion (SIC), and Akaike information criterion (AIC)) [41]–[43], we obtain a standard performance appraisal through this work. To find the best performing model, we test various input parameters, p, q, P and Q range from 1 to 5, d and D range from 1 to 2, with seasonal period s as 30. We choose the optimal parameters by grid search approach and select the model yielding the lowest MAPE in the testing sample forecast as the best one.

3) NNAR MODEL

ANN, a powerful ML method, has been extensively used and applied in many domains, ranging from data science to natural language processing, and computer vision. A Feed-forward Neural Network (FNN) is one type of ANN with one or more than one hidden layer and several lagged inputs. The architecture of an FNN is normally exhibited as Fig. 9 − the simplest kind of FNN, consisting of a single hidden layer. In FNN, each layer contains several nodes and there are several connections among nodes which are from a specific layer to the next layer. However, the nodes in a given layer are mutual independent (i.e., no connections among different nodes in the same layer).

With time series, we put lagged data as inputs to a neural network, this is the same as we applied lagged inputs in SARIMA model in Section III-D2. We only consider single layer FNN, because theoretically, a neural network with single hidden layer and multiple hidden neurons can approximate any continuous function [44], and with the increase

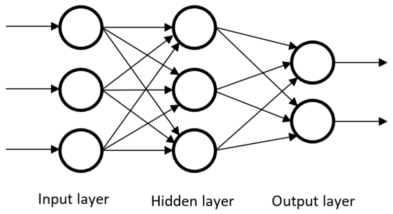


FIGURE 9. FNN with three inputs and one hidden layer with three hidden neurons.

of hidden layers, the computation time increases. We train a Neural Network Auto-Regression model called NNAR $(p, P, k)_s$ model, where p is the lagged input, k represents the number of nodes in the hidden layer, P refers to the order of seasonal AR, and s shows the seasonal period.

Consistent with the model selection procedure mentioned above, we vary the values of each parameter to achieve the best model that performs the lowest MAPE over the testing period. The input parameters p and P range from 1 to 20, k ranges from 1 to 5, and seasonal period s is 30.

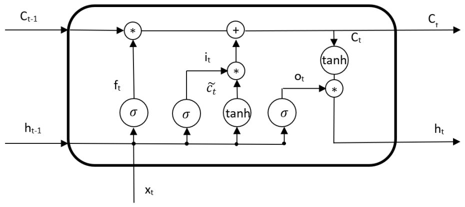


FIGURE 10. LSTM memory cell topology of the hidden layer.

4) LSTM MODEL

Hochireiter and Schmidhuber introduced LSTM in 1997 [45] to overcome the vanishing and exploding gradients of Recurrent Neural Networks (RNN). Although, RNN can deal with sequences of inputs, such as time series, text and audio, RNN has an inherent problem with long data and made the model less sensitive with longer input data. LSTM model can avoid this problem thanks to a more complex memory cell. Fig. 10 illustrates the cell structure which is the key design of LSTM model. Three particular gates are designed including a forgot gate, an input gate and an output gate. Different from ANN, the hidden layers of LSTM network are connected with each other. The inputs of hidden layers include not only the input of input layer but also the output of hidden layer. LSTM model has the ability to remember information of long-term, and it is a desirable approach for long-term dependency tasks.

The information processing mechanism of an LSTM model is described as:

$$f_t = \sigma(W_f x_f + R_f h_{t-1} + b_f), \tag{16}$$

$$i_t = \sigma(W_i x_t + R_i h_{t-1} + b_i), \tag{17}$$

$$o_t = \sigma(W_o x_t + R_o h_{t-1} + b_o), \tag{18}$$

$$\widetilde{c_t} = \tanh(W_c x_t + R_c h_{t-1} + b_c), \tag{19}$$

$$c_t = f_t * c_{t-1} + i_t * \widetilde{c_t}, \tag{20}$$

$$h_t = o_t * \tanh c_t, \tag{21}$$

where the notations are defined in Table 2.

TABLE 2. Parameters and variables in LSTM model.

Parameters and variables	Meaning
σ	Sigmoid function, range from 0 to 1
$\tanh$	Hyperbolic tangent function, range from -1 to 1
$*$	Element-wise multiplication
i_t	Input gate
f_t	Forgot gate
o_t	Output gate
c_t	State of current memory cell at time t
$\widetilde{c_t}$	Candidate value for state at time t
h_t	Output value
x_t	Input value
W_i, W_f, W_o, W_c	Weights
R_i, R_f, R_o, R_c	Weights
b_i, b_f, b_o, b_c	Bias vectors

In order to find the best-performing LSTM model, the range of hidden neurons is set to 1 to 400, time step is set to 1 to 20, and batch size is set to 2 to 64. Meanwhile, we adjust the optimizer function, activation function and loss function. Candidate models with multiple parameter combinations are iterated through 100 epochs. To avoid over-fitting problem, an Early Stopping function is applied. Fig. 11 indicates the training process at $x = 150m$. The plotted curve decreases and remains stable before 100-step iteration, therefore, no over-fitting is observed.

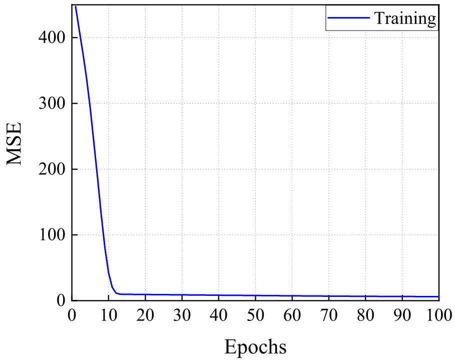


FIGURE 11. Training performance for LSTM model at 100 epochs at $x = 150m$.

To sum up, we introduce four forecasting models in Section III-D. In EOF model, parameters are determined by decomposition and linear regression. In the other models, we test a variety of combinations of input parameters to find the best performing models with grid search approach. The results of single-step and multi-step predictions are reported in the next section.

IV. RESULTS AND DISCUSSIONS

This section first determines the best imputation method to obtain continuous time series for training the models. Then, we evaluate forecasting performance of various models for

TABLE 3. MAPE of forecasting models based on different imputation methods (The bold values represent the best MAPE for each model).

Models	Imputation approaches	Positions					
		$50m$	$100m$	$150m$	$200m$	$250m$	Average
SARIMA	Zero imputation	9.717%	8.298%	11.121%	9.531%	8.270%	9.387%
	Seasonal adjusted LOCF	8.487%	6.717%	5.394%	**3.344%**	2.520%	5.292%
	Seasonal adjusted linear interpolation	**8.368%**	**6.549%**	**5.309%**	3.359%	**2.494%**	**5.216%**
	KNN	9.821%	8.280%	6.167%	4.600%	3.422%	6.458%
NNAR	Zero imputation	10.788%	9.646%	9.090%	7.765%	6.094%	8.677%
	Seasonal adjusted LOCF	**7.893%**	6.492%	4.908%	3.272%	2.619%	5.037%
	Seasonal adjusted linear interpolation	7.929%	**6.459%**	**4.864%**	**3.139%**	**2.483%**	**4.975%**
	KNN	9.650%	8.922%	6.164%	4.896%	3.862%	6.699%
LSTM	Zero imputation	11.856%	9.403%	7.843%	4.657%	3.505%	7.453%
	Seasonal adjusted LOCF	8.369%	6.115%	**4.685%**	**3.378%**	2.456%	5.000%
	Seasonal adjusted linear interpolation	**8.282%**	**6.087%**	4.755%	3.398%	**2.439%**	**4.992%**
	KNN	8.932%	6.703%	4.985%	3.424%	2.440%	5.297%

single-step prediction (one day ahead) and multi-step prediction (up to 50 days ahead). The predictions of each model are compared to the realities based on four performance metrics documented in Section III-B. Furthermore, we assess the predictive ability of various model through line plots. Lastly, Taylor diagrams are presented to summarize different performance metrics.

A. MISSING DATA IMPUTATION

As stated in Section III-A, we introduce four missing data imputation methods, zero imputation, seasonal adjusted LOCF, seasonal adjusted linear interpolation and KNN. These four methods have been extensively used by previous literature [34]–[39].

We select the best imputation approach based on the following process. First, after filling missing values via the candidate imputation methods, the time series at each position now has continued training samples which are available to train SARIMA, NNAR and LSTM model. Then, we measure the MAPE for these models at each position in Table 3. Lastly, the filling method yielding the lowest averaged MAPE across five positions is selected as the best.

Across four filling methods, Table 3 shows that zero imputation method yields the highest MAPE. As this basic method fills the missing values with zero, which is not consistent with the nature of series, e.g., the cross-shore distance cannot be zero in any scenario. Hence, this approach brings a large amount of bias information into the imputed series. Meanwhile, the seasonality of the original series is weakened by zero imputation as such a method fails to account for any seasonal features.

Seasonal adjusted LOCF and seasonal adjusted linear interpolation method produce better performance than zero imputation. More specific, linear interpolation outperforms LOCF. For SARIMA model, linear interpolation method generates the lowest MAPE at most positions except 200m, and the average MAPE is the lowest, in comparison with other filling methods. For NNAR and LSTM models, linear interpolation method is also the most desirable imputation approach. It achieves the lowest MAPE at most positions, and the average MAPE is the lowest across four methods. This shows that seasonal adjustment is very effective when seasonality is obvious [46].

KNN shows better performance than zero imputation but weaker than seasonal adjusted linear interpolation. The average MAPE of SARIMA with KNN is around 3 % lower than that with zero imputation, and the average MAPE of NNAR and LSTM with KNN model is around 2 % lower than the MAPE produced by zero imputation.

In our study, missing values are imputed by seasonal adjusted linear interpolation approach. After imputation, the time series is available for training the models and implementing the predictions.

B. SINGLE-STEP PREDICTION

Various statistical performance metrics including R, MAE, RMSE and MAPE are applied to make comparisons. Table 4 summarizes forecasting performance of different models, where real data is used as the benchmark. The optimal parameters for SARIMA, NNAR, and LSTM are reported in Table 5, 6, and 7, respectively.

Table 4 shows that EOF model is less accurate than other models in terms of statistical performance. The MAPE of this model is the highest at each location. The worst MAPE occurs at $x = 50m$ as 17.522%. Although EOF model reports R greater than 0.9 at $50m$ and $100m$ as 0.964 and 0.935, respectively, R reduces to below 0.9 at all remaining positions whereas the other models do not. Therefore, conventional EOF model is not a desirable forecasting approach that works for every case.

SARIMA model outperforms EOF model in terms of statistical performance. Although it generates the second highest MAPE at all positions except $200m$, the superiority is obvious on average, 6.327%, where the greatest difference is 9.154% at $x = 50m$. Meanwhile, this model yields a higher R than EOF model at each position, where the largest difference is 0.086 at $x = 200m$. As for MAE and RMSE, this model yields much better performance than EOF model regardless of the

TABLE 4. Performance metrics of four models for single-step prediction at each position (The bold values represent the best performance for each metric).

Models	Performance Metrics	Positions				
		$50m$	$100m$	$150m$	$200m$	$250m$
EOF	R	0.964	0.935	0.897	0.869	0.874
	RMSE	4.904	3.719	3.464	3.607	3.756
	MAE	4.004	3.028	2.803	2.884	3.106
	MAPE	17.522%	14.006%	11.090%	8.133%	6.965%
SARIMA	R	0.979	0.974	0.964	0.955	0.953
	RMSE	2.077	1.688	1.696	1.700	1.503
	MAE	1.493	1.208	1.145	1.071	1.077
	MAPE	8.368%	6.549%	5.309%	3.359%	2.494%
NNAR	R	**0.981**	**0.976**	**0.969**	**0.964**	**0.956**
	RMSE	**2.021**	**1.664**	1.582	**1.576**	1.480
	MAE	**1.430**	1.228	1.105	**1.049**	1.082
	MAPE	**7.929%**	6.459%	4.864%	**3.139%**	2.483%
LSTM	R	0.979	0.975	**0.969**	0.959	0.955
	RMSE	2.075	1.667	**1.574**	1.628	**1.469**
	MAE	1.485	**1.159**	**1.077**	1.096	**1.068**
	MAPE	8.282%	**6.087%**	**4.755%**	3.398%	**2.439%**

TABLE 5. The optimal parameters of SARIMA.

Position	SARIMA $(p, d, q) \times (P, D, Q)_s$
50m	$(0, 1, 0) \times (1, 0, 0)_{30}$
100m	$(0, 1, 0) \times (3, 0, 0)_{30}$
150m	$(2, 1, 0) \times (1, 0, 1)_{30}$
200m	$(0, 1, 0) \times (1, 0, 1)_{30}$
250m	$(0, 1, 0) \times (2, 0, 0)_{30}$

TABLE 6. The optimal parameters of NNAR model.

Position	NNAR $(p, P, k)_m$
50m	$(3, 8, 3)_{30}$
100m	$(1, 1, 4)_{30}$
150m	$(3, 2, 1)_{30}$
200m	$(13, 1, 1)_{30}$
250m	$(1, 8, 1)_{30}$

TABLE 7. The optimal parameters of LSTM network topology.

Parameters	$50m$	$100m$	$150m$	$200m$	$250m$
epoch			100		
batch size	2	8	16	8	4
time step			1		
Unit	250	300	300	300	150
Activation function			Relu		
Optimizer	Adamax	Adam	Adadelta	Nadam	Adadelta
loss function			MSE		

positions. Therefore, SARIMA model is a more desirable model than EOF model.

Furthermore, NNAR model outperforms EOF model in terms of different metrics. This model reports the second lowest MAPE at $x = 100m$, $150m$ and $250m$ and the lowest value at remaining positions. The MAPE of NNAR is less than half of that of EOF at all positions, and around 0.45% lower than that of SARIMA at $x = 50m$ and $x = 150m$. Meanwhile, it yields higher values of R than the above-mentioned models at each position. As for the other two metrics, NNAR model has better performance in RMSE than SARIMA and EOF model at all positions, and it also yields lower MAE than SARIMA and EOF model at $x = 50m$, $150m$ and $200m$. This implies that NNAR model is also a more desirable model than EOF model.

LSTM displays similar forecasting ability compared to NNAR model. In terms of MAPE and MAE, LSTM model outperforms other models at $x = 100m$, $150m$ and $250m$, and second lowest values at $x = 50m$. Similarly, it achieves the highest R, 0.969, at $x = 150m$, but slightly lower than that of NNAR at the remaining positions. For RMSE, this model displays the best performance at $x = 150m$ and $250m$ and the second best performance at the remaining points.

We further focus on line plots of single-step predictions of EOF, SARIMA, NNAR and LSTM at $x = 50m$. The left panel of Fig. 12 plots the forecasting curves, in which the predicted values exhibit the same trend as realities regardless of models. This implies that every involved model can capture seasonality and patterns, hence, these models are worth to investigate. More specifically, EOF model presents the poorest predictive ability as it only captures the trend but fails to predict the shocks. By contrast, the predictive abilities of SARIMA, NNAR and LSTM model are superior to that of EOF model. These three models can predict not only the trend but also the shocks during the testing period. The prediction curves of SARIMA, NNAR and LSTM models are almost identical. The right panel of Fig. 12 displays the correlation between prediction and realities, which is equivalent to R in Table 4. In line with our previous finding, the R of SARIMA, NNAR, and LSTM are higher than that of EOF.

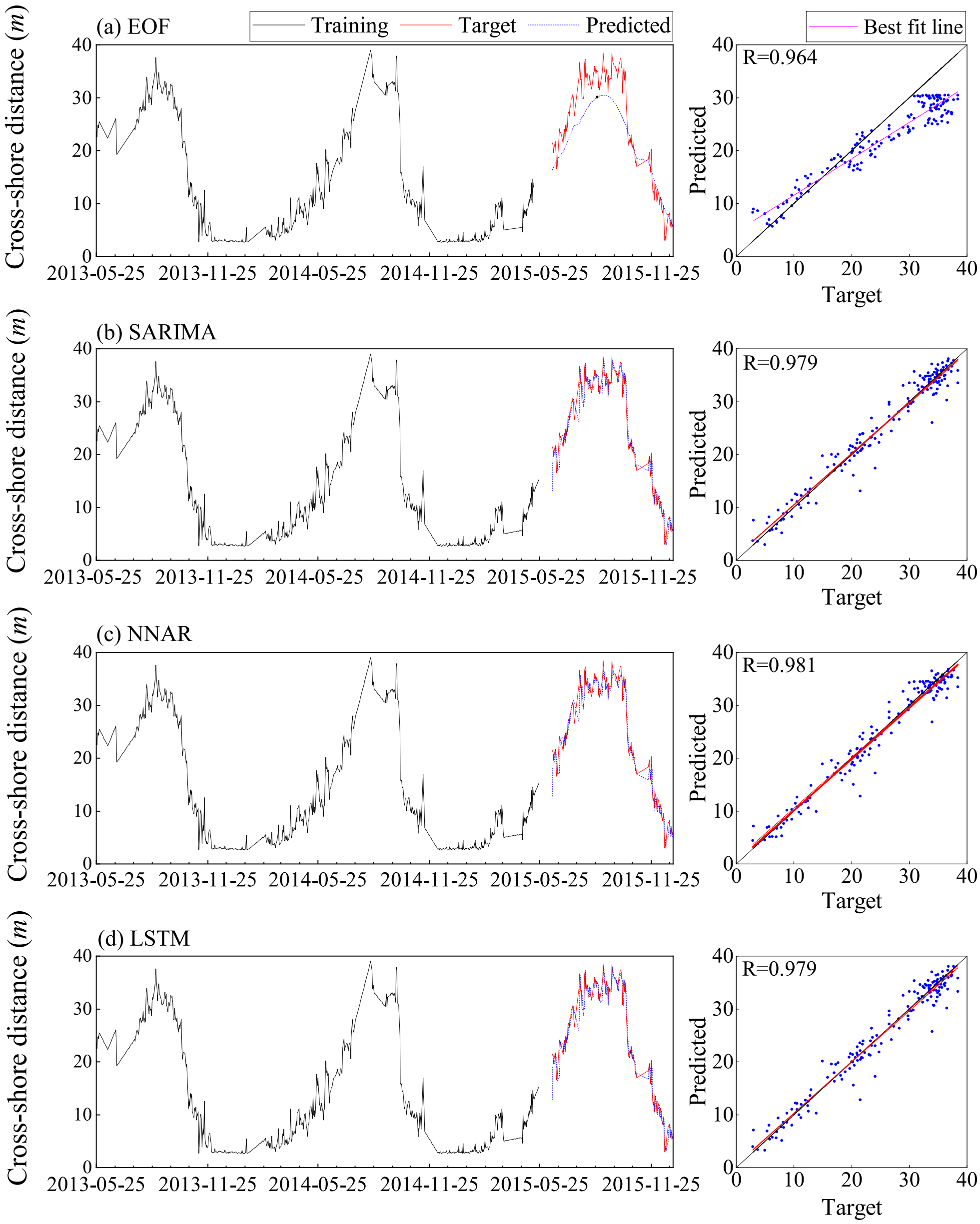


FIGURE 12. Forecasting performance of four models for single-step prediction at $x = 50m$.

The predictions at the remaining positions also display similar patterns. We plot the forecasting curves at each position in Fig. 13, where the real values are used as the benchmark. In each sub-figure, all four models display the same trend as the real values. Consistent with the results in Table 4, EOF shows the poorest predictive ability and fails to forecast any shocks, whereas SARIMA, NNAR and LSTM predict both the trends and shocks over the testing period. Overall, the forecasting performance of SARIMA, NNAR and LSTM is similar.

We further present Taylor diagrams to compare the performance of four models in terms of Pearson Correlation Coefficient R, centred Root-Mean-Square Difference RMSD_c and Standard Deviation SD [47]. The definition of R is presented in (4). SD is used to describe the variation of series. To be specific, a low SD demonstrates that the values are close to the mean value. The definitions of SD and RMSD_c [47] are given as:

$$\text{SD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(B_i - \bar{B})^2}, \tag{22}$$

$$\text{RMSD}_\text{c} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[(A_i - \bar{A})] - (B_i - \bar{B})]^2}, \tag{23}$$

where n is the number of observations, A_i is the real value, B_i is the prediction, $\bar{A}$ and $\bar{B}$ are the mean of these two series.

As shown in Fig. 14, EOF model has the poorest performance as it locates the farthest from the Target point.

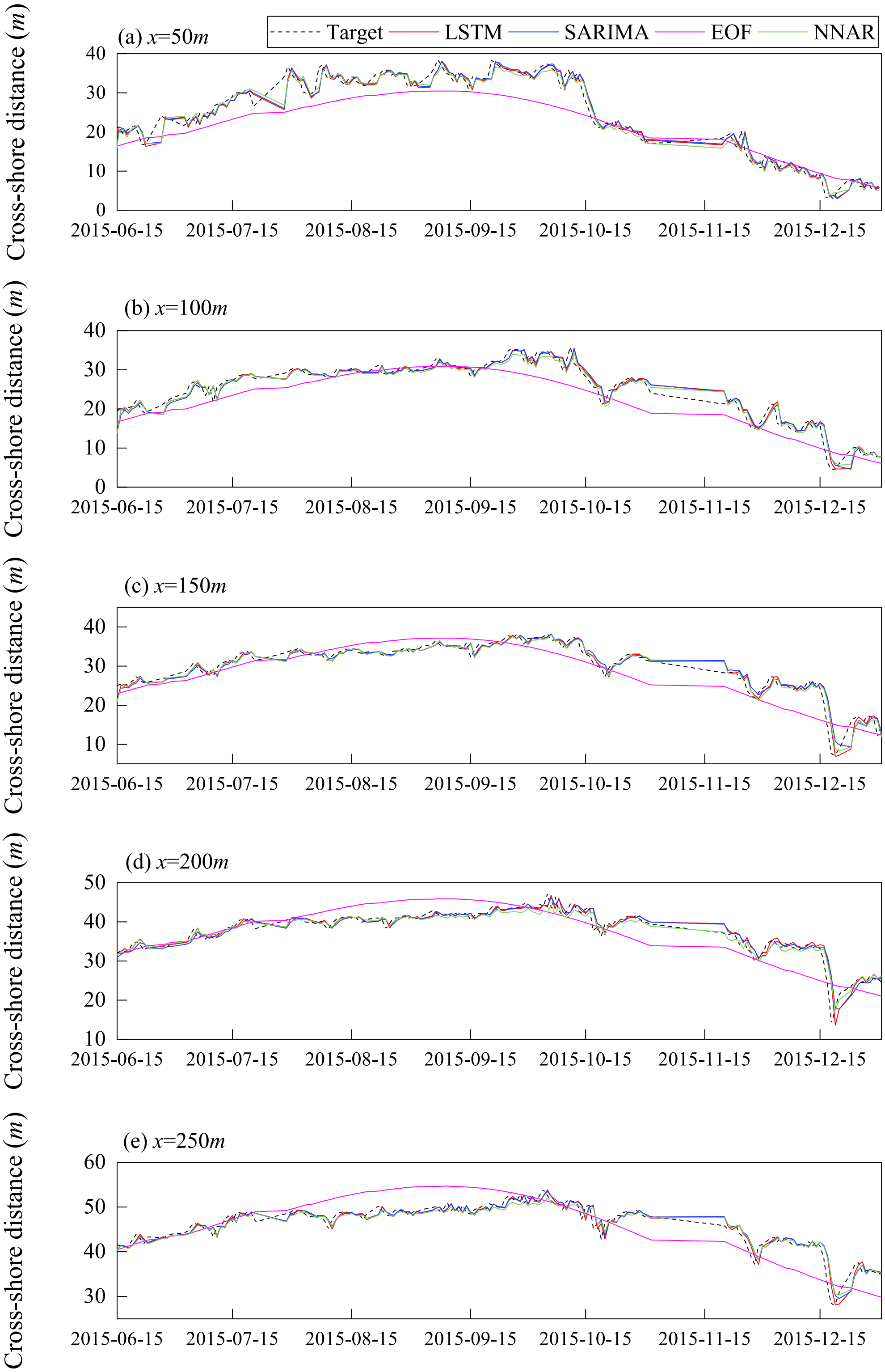


FIGURE 13. Forecasting performance of four model for single-step prediction.

This is consistent with our findings in Table 4, Fig. 12 and Fig. 13. The performance differences among SARIMA, NNAR and LSTM are not significant as they almost overlap at $x = 50m$, $x = 100m$ and $150m$. At $x = 200m$, NNAR locates closer to the Target than SARIMA and LSTM, but the superiorities are still tiny.[2] On average, the differences among SARIMA, NNAR and LSTM are not

[2] The Taylor diagram at $x = 250m$ shows similar patterns to other positions.

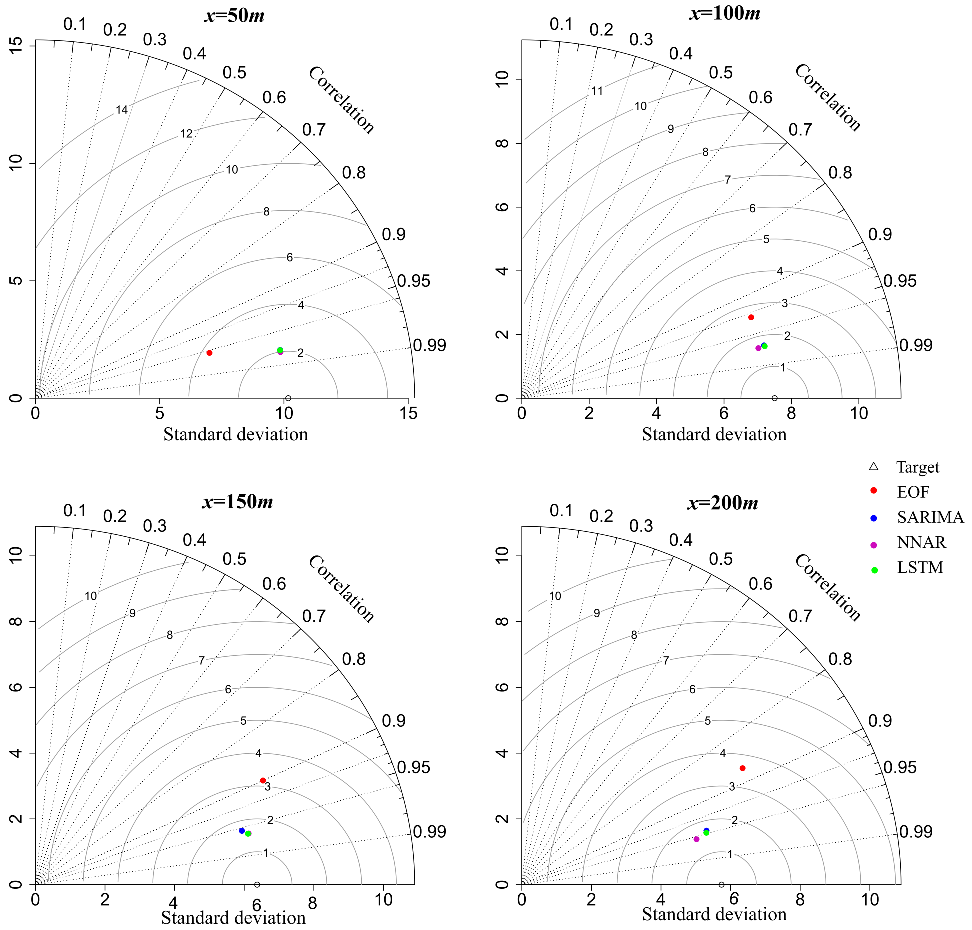


FIGURE 14. Taylor diagrams to compare the performance of four models for single-step prediction.

distinct in Taylor diagrams as their positions are close to each other.

This section summarizes the following highlights. First, EOF model shows the poorest predictive ability at each location regardless of performance metrics. EOF model follows assumptions that the temporal eigenfunction of shoreline variation can be decomposed by (13). In this equation, the first component is used to represent the entire shoreline variation and the other components are ignored. This is a strict assumption and may not fit all data, so the prediction curve of EOF is smooth and fails to predict shocks, as shown in Fig. 12 and Fig. 13. Second, statistical model SARIMA outperforms EOF model and is proved to be a reliable forecasting method. ML models NNAR and LSTM exhibit similar predictive abilities to SARIMA on average. Different from EOF model, these three models are based on data characteristics rather than physical theorems, so they are able to predict the trends as well as the shocks. Therefore, to achieve accurate results for single-step prediction, SARIMA, NNAR and LSTM are more desirable models than EOF for forecasting shoreline changes.

C. MULTI-STEP PREDICTION

While most existing forecasting methods only focus on single-step prediction, we are the first to forecast the shoreline changes in Nha Trang Coast very far into the future. In this section, we study the performance of up to 50-day ahead prediction to see how our models perform in a long term. The parameters in each model are consistent with single-step prediction as determined in Section IV-B, and multi-step predictions are evaluated by using the data at $x = 50m$. The forecasting step K ranges from 10 to 50 with an interval of 10 steps. Table 8 reports the long-term forecasting abilities among SARIMA, NNAR and LSTM models. Compared to single-step prediction, multi-step prediction shows worse results due to the cumulative errors discussed in Section III-C.

Table 8 shows that SARIMA model presents a decreasing performance ability with K increasing from 1 to 50. The performance of ML models exhibits similar changes to that of SARIMA model. The forecasting ability of both NNAR and LSTM reduces with the increase of forecasting steps. NNAR displays the best predictive ability in terms of four

TABLE 8. **Performance metrics of models for multi-step prediction at x = 50m (The bold values represent the best performance for each metric).**

Models	performance metric	Steps					
		K=1	K=10	K=20	K=30	K=40	K=50
SARIMA	R	0.979	0.875	0.726	0.622	0.422	0.101
	RMSE	2.077	4.976	7.179	8.334	10.133	12.693
	MAE	1.493	3.950	5.832	6.880	8.827	11.146
	MAPE	8.368%	23.844%	36.031%	45.593%	56.709%	72.603%
NNAR	R	**0.981**	**0.894**	**0.761**	**0.662**	**0.474**	**0.163**
	RMSE	**2.021**	**4.588**	**6.729**	**7.931**	**9.731**	12.275
	MAE	**1.430**	**3.627**	**5.423**	**6.461**	**8.427**	**10.715**
	MAPE	**7.929%**	**21.445%**	**33.041%**	**41.998%**	**53.021%**	68.705%
LSTM	R	0.979	0.876	0.728	0.629	0.418	0.088
	RMSE	2.075	5.029	7.269	8.467	10.388	**12.224**
	MAE	1.485	4.179	6.276	7.815	9.818	11.239
	MAPE	8.282%	23.499 %	34.350 %	46.210%	54.970%	**63.004%**

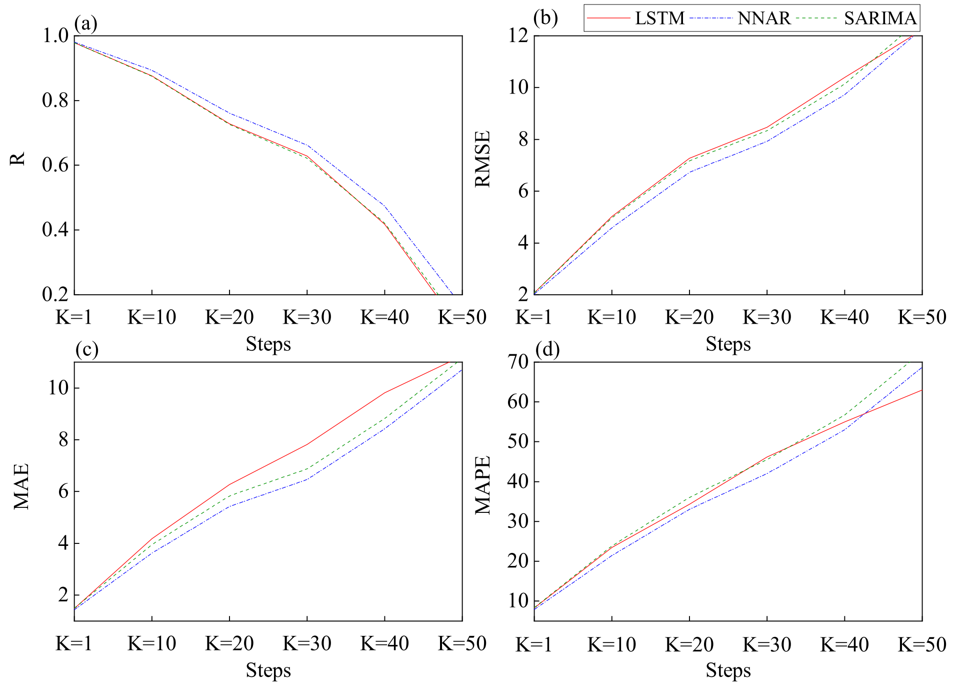


FIGURE 15. **Comparisons of performance metrics for multi-step prediction at $x = 50m$.**

metrics from $K = 10$ to $K = 40$. It yields the highest R, the lowest RMSE, MAE, and the lowest MAPE across three models. Differently, at $K = 50$, LSTM model reports the lowest RMSE and MAPE, but its R is still lower than NNAR, and its MAE exceeds that of NNAR. Overall, for long-term forecasting, NNAR is slightly better than the other two models, despite the superiorities are not significant in every performance metric. This further supports our above-mentioned conclusion that the predictive ability of SARIMA is comparable to that of the ML models as discussed in Section IV-B.

To display the dynamic of predictive ability for multi-step prediction of each model, Fig. 15 plots the changes of performance metrics with the increase of prediction steps K. The forecasting performance decreases gradually, not suddenly at some points. That shows the robustness of forecasting models. The line charts further support that the differences among SARIMA, NNAR, and LSTM are not significant. The R of SARIMA and LSTM shows larger reduction than that of NNAR. The RSME and MAE of NNAR model increase less than those of SARIMA and LSTM models. The MAPEs of three models are almost identical until $K = 40$, but the MAPEs of SARIMA and NNAR are higher than that of LSTM at $K = 50$. Those patterns imply that NNAR yields smaller prediction errors in the testing periods despite the differences are tiny on average.

Section IV compares the performance of EOF, SARIMA, NNAR and LSTM in terms of single-step and multi-step forecasting. Multi-step prediction shows worse performance due to the cumulative errors. However, the forecasting performance when implementing 10-day ahead prediction is still desirable (R > 0.875 for all models). When there is a disaster (e.g., storm), a one-day early warning can help to execute many necessary plans to avoid critical damages to community. Despite the multi-step predictions exhibit the long-term predictive abilities for our models, the accuracy for a few days ahead prediction is the most important. In addition, our results also indicate that SARIMA, NNAR and LSTM present similar forecasting abilities and significantly outperform conventional EOF model studied in [4]. Therefore, statistical forecasting methods, e.g., SARIMA model, are still desirable approaches for forecasting tasks, which is in line with previous studies [17]–[19]. Consistent with findings in [21], [29], sophisticated ML models do not always outperform statistical methods although ML methods are associated with high computational complexity.

We further discuss the potential reasons that ML methods, especially deep learning LSTM model, do not show advanced results. As stated in [29], the characteristic and the length of specific time series are important factors that could potentially influence the accuracy of various forecasting models. The performance of different forecasting models depends on the characteristic of data, and ML techniques could perform better with high complexity and strong nonlinear time series. In addition, ML methods might lead to better results than statistical methods when the time series data is sufficient as their parameters can be trained optimally, otherwise, proper training is hard to be realized with short time series. In this research, LSTM model might show greater performance if more observations are available. This needs to be verified by more experiments with various lengths of training series for future research. Our findings highlight the importance of evaluating not only complicated ML models but also statistical models for forecasting tasks.

V. CONCLUSION

This paper explores four data imputation methods, zero imputation, seasonal adjusted linear interpolation, seasonal adjusted LOCF, and KNN, to provide quality coastal data for feeding the prediction models. The results show that seasonal adjusted linear interpolation method is the most desirable imputation approach. More importantly, this work further investigates reliable models to predict shoreline variations. Statistical forecasting model SARIMA and ML models, NNAR and LSTM are applied to perform single-step and multi-step predictions (up to 50 days ahead). By comparing different performance metrics, plots and Taylor diagrams, our results demonstrate that SARIMA, NNAR and LSTM outperform EOF significantly, and all these three models can be effectively used for monitoring coastal variations from video cameras under extreme weather conditions.

ACKNOWLEDGEMENT

This work was supported by a Researcher Links grant, ID 527612186, under the Newton Fund partnership. The grant is funded by the UK Department for Business, Energy and Industrial Strategy and delivered by the British Council, for further information, please visit www.newtonfund.ac.uk.

REFERENCES

[1] E. C. Bird, *Beach Management*, vol. 5. New York, NY, USA: Wiley, 1996.

[2] T. M. Thanh, Y. Mitobe, V. C. Hoang, N. T. Viet, and H. Tanaka, "Coastal morphology change and its relationship with climate characteristics on Nha Trang Coast, Vietnam," in *Proc. ICEC*, Muscat, Oman, 2015, pp. 159–164.

[3] X. T. Nguyen, M. T. Tran, H. Tanaka, T. V. Nguyen, Y. Mitobe, and C. D. Duong, "Numerical investigation of the effect of seasonal variations of depth-of-closure on shoreline evolution," *Int. J. Sediment Res.*, vol. 36, no. 1, pp. 1–16, Feb. 2021.

[4] T. M. Thanh, H. Tanaka, Y. Mitobe, N. T. Viet, and R. Almar, "Seasonal variation of morphology and sediment movement on Nha Trang Coast, Vietnam," *J. Coastal Res.*, vol. 81, no. 1, pp. 22–31, Sep. 2018.

[5] S. A. Hughes, *Physical Models and Laboratory Techniques in Coastal Engineering*, vol. 7. Singapore: World Sci., 1993.

[6] S. Zeinali, M. Dehghani, and N. Talebbeydokhti, "Artificial neural network for the prediction of shoreline changes in Narrabeen, Australia," *Appl. Ocean Res.*, vol. 107, Feb. 2021, Art. no. 102362.

[7] P. Ekphisutsuntorn, P. Wongwises, C. Chinnarasri, U. Humphries, and S. Vongvisessomjai, "Numerical modeling of erosion for muddy coast at Bangkhuntien shoreline, Thailand," *Int. J. Environ. Sci. Technol.*, vol. 2, no. 4, pp. 230–240, 2010.

[8] B. C. Douglas and M. Crowell, "Long-term shoreline position prediction and error propagation," *J. Coast. Res.*, vol. 16, no. 1, pp. 145–152, Dec. 2000.

[9] A. Santra, D. Mitra, and S. Mitra, "Spatial modeling using high resolution image for future shoreline prediction along Junput Coast, West Bengal, India," *Geo-Spatial Inf. Sci.*, vol. 14, no. 3, pp. 157–163, Jan. 2011.

[10] E. N. Lorenz, *Empirical Orthogonal Functions and Statistical Weather Prediction*. Cambridge, MA, USA: Massachusetts Institute of Technology, Department of Meteorology, Dec. 1956.

[11] S. Sharma, K. Hamal, N. Khadka, and B. B. Joshi, "Dominant pattern of year-to-year variability of summer precipitation in nepal during 1987–2015," *Theor. Appl. Climatol.*, vol. 142, nos. 3–4, pp. 1071–1084, Aug. 2020.

[12] B. Xiang, Y. Q. Sun, J. Chen, N. C. Johnson, and X. Jiang, "Subseasonal prediction of land cold extremes in boreal wintertime," *J. Geophys. Res., Atmos.*, vol. 125, no. 13, p. 32670, Jul. 2020.

[13] P. Liang, H. Lin, and Y. Ding, "Dominant modes of subseasonal variability of east asian summertime surface air temperature and their predictions," *J. Climate*, vol. 31, no. 7, pp. 2729–2743, Apr. 2018.

[14] A. Wiese, J. Staneva, J. Schulz-Stellenfleth, A. Behrens, L. Fenoglio-Marc, and J.-R. Bidlot, "Synergy of wind wave model simulations and satellite observations during extreme events," *Ocean Sci.*, vol. 14, no. 6, pp. 1503–1521, Dec. 2018.

[15] P. Whittle, "Tests of fit in time series," *Biometrika*, vol. 39, nos. 3–4, pp. 309–318, 1952.

[16] A. I. McKerchar and J. W. Delleur, "Application of seasonal parametric linear stochastic models to monthly flow data," *Water Resour. Res.*, vol. 10, no. 2, pp. 246–255, Apr. 1974.

[17] Q. Sun, J. Wan, and S. Liu, "Estimation of sea level variability in the China Sea and its vicinity using the SARIMA and LSTM models," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3317–3326, Jun. 2020.

[18] D. B. Alencar, C. M. Affonso, R. C. L. Oliveira, and J. C. R. Filho, "Hybrid approach combining SARIMA and neural networks for multi-step ahead wind speed forecasting in Brazil," *IEEE Access*, vol. 6, pp. 55986–55994, Oct. 2018.

[19] S. Yang, Z. Zhang, L. Fan, T. Xia, S. Duan, C. Zheng, X. Li, and H. Li, "Long-term prediction of significant wave height based on SARIMA model in the South China Sea and adjacent waters," *IEEE Access*, vol. 7, pp. 88082–88092, Jun. 2019.

[20] A. Maleki, S. Nasseri, M. S. Aminabad, and M. Hadi, "Comparison of ARIMA and NNAR models for forecasting water treatment plant's influent characteristics," *KSCE J. Civil Eng.*, vol. 22, no. 9, pp. 3233–3245, Apr. 2018.

[21] R. Thoplan, "Simple v/s sophisticated methods of forecasting for mauritius monthly tourist arrival data," *Int. J. Stats. Appl.*, vol. 4, no. 5, pp. 217–223, Apr. 2014.

[22] E. S. Silva, H. Hassani, S. Heravi, and X. Huang, "Forecasting tourism demand with denoised neural networks," *Ann. Tourism Res.*, vol. 74, pp. 134–154, Jan. 2019.

[23] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[24] S. Biswas and M. Sinha, "Performances of deep learning models for Indian Ocean wind speed prediction," *Model. Earth Syst. Environ.*, vol. 7, pp. 1–23, Sep. 2020.

[25] E.-J. Lee, J.-Y. Chae, and J.-H. Park, "Reconstruction of sea level data around the Korean Coast using Artificial neural network methods," *J. Coastal Res.*, vol. 95, pp. 1172–1176, May 2020.

[26] S. Fan, N. Xiao, and S. Dong, "A novel model to predict significant wave height based on long short-term memory network," *Ocean Eng.*, vol. 205, Jun. 2020, Art. no. 107298.

[27] Y. Wang, C. Xu, S. Zhang, L. Yang, Z. Wang, Y. Zhu, and J. Yuan, "Development and evaluation of a deep learning approach for modeling seasonality and trends in hand-foot-mouth disease incidence in mainland China," *Sci. Rep.*, vol. 9, no. 1, pp. 1–15, May 2019.

[28] Z. Pala and R. Atici, "Forecasting sunspot time series using deep learning methods," *Sol. Phys.*, vol. 294, no. 5, p. 50, May 2019.

[29] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and machine learning forecasting methods: Concerns and ways forward," *PLoS ONE*, vol. 13, no. 3, Mar. 2018, Art. no. 0194889

[30] N. T. Viet, N. V. Duc, V. Hoang, H. Tanaka, T. Tung, J. Lefebvre, and R. Almar, "Investigation of erosion mechanism on Nha Trang Coast, Vietnam," in *Proc. 19th IAHR-APD*, Hanoi, Vietnam, 2014, pp. 1–7.

[31] K. T. Holland, R. A. Holman, T. C. Lippmann, J. Stanley, and N. Plant, "Practical use of video imagery in nearshore oceanographic field studies," *IEEE J. Ocean. Eng.*, vol. 22, no. 1, pp. 81–92, Jan. 1997.

[32] E. H. Boak and I. L. Turner, "Shoreline definition and detection: A review," *J. Coastal Res.*, vol. 214, pp. 688–703, Jul. 2005.

[33] V. Phan, T. Ngo-Duc, and T. Ho, "Seasonal and interannual variations of surface climate elements over Vietnam," *Climate Res.*, vol. 40, pp. 49–60, Sep. 2009.

[34] R. Kabacoff, *R in Action*. Shelter Island, NY, USA: Manning Publ., 2011.

[35] S. J. Choudhury and N. R. Pal, "Imputation of missing data with neural networks for classification," *Knowl.-Based Syst.*, vol. 182, Oct. 2019, Art. no. 104838.

[36] R. H. Shumway, D. S. Stoffer, and D. S. Stoffer, *Time Series Analysis and its Applications*, vol. 3. New York, NY, USA: Springer, 2000.

[37] C. K. Enders, *Applied Missing Data Analysis*. New York, NY, USA: Guilford Press, 2010.

[38] E. Meijering, "A chronology of interpolation: From ancient astronomy to modern signal and image processing," *Proc. IEEE*, vol. 90, no. 3, pp. 319–342, Mar. 2002.

[39] G. E. Batista and M. C. Monard, "A study of K-nearest neighbour as an imputation method," *HIS*, vol. 87, nos. 251–260, p. 48, Dec. 2002.

[40] V. M. Landassuri-Moreno and J. A. Bullinaria, "Neural network ensembles for time series forecasting," in *Proc. GECCO*, Montreal, QC, Canada, 2009, pp. 1235–1242.

[41] M. Zeynoddin, H. Bonakdari, I. Ebtehaj, F. Esmaeilbeiki, B. Gharabaghi, and D. Z. Haghi, "A reliable linear stochastic daily soil temperature forecast model," *Soil Tillage Res.*, vol. 189, pp. 73–87, Jun. 2019.

[42] I. Ebtehaj, H. Bonakdari, and B. Gharabaghi, "A reliable linear method for modeling lake level fluctuations," *J. Hydrol.*, vol. 570, pp. 236–250, Mar. 2019.

[43] H. Bonakdari, H. Moeeni, I. Ebtehaj, M. Zeynoddin, A. Mahoammadian, and B. Gharabaghi, "New insights into soil temperature time series modeling: Linear or nonlinear?" *Theor. Appl. Climatol.*, vol. 135, nos. 3–4, pp. 1157–1177, Mar. 2018.

[44] L. F. Guilhoto. (2018). *An Overview of Artificial Neural Networks for Mathematicians*. [Online]. Available: https://math.uchicago.edu/ may/REU2018/REUPapers/Guilhoto.pdf

[45] S. Hochreiter and J. Schmidhuber, "Lstm can solve hard long time lag problems," in *Proc. Adv. Neural Inf. Process. Syst.*, 1997, pp. 473–479.

[46] Y. Qin, G. Ren, P. Zhang, L. Wu, and K. Wen, "An imputation method for the climatic data with strong seasonality and spatial correlation," *Theor. Appl. Climatol.*, vol. 144, nos. 1–2, pp. 203–213, Jan. 2021.

[47] K. E. Taylor, "Summarizing multiple aspects of model performance in a single diagram," *J. Geophys. Res., Atmos.*, vol. 106, no. D7, pp. 7183–7192, Apr. 2001.

CHENG YIN received the Ph.D. degree in wireless communication from Queen's University Belfast (QUB), U.K., in 2019. She is currently a Research Fellow with the Centre for Wireless Innovation (CWI), QUB. Her research interests include machine learning, wireless communications, physical layer security, and green communication networking.

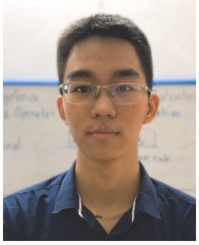
LE THANH BINH received the Ph.D. degree in hydraulic construction engineering from Thuyloi University, in 2017. He is a Technical Assistant of the General Director of Vietnam Hydraulic Engineering Consultants Corporation—JSC (HEC), with respect to marine construction. The company is the leading unit in the field of marine construction of ministry of agriculture and rural development in Vietnam. He is a senior in bathymetry, hydrological, and shoreline monitoring system by video-camera monitoring technology. He is currently the Key Member of the team in charge of installing, maintaining, analyzing, and evaluating all coastal shore-based camera systems in central Vietnam. His current research interests include studying shoreline change and sediment transport.

DUONG TRAN ANH received the master's degree from the Water Engineering And Management, Asian Institute of Technology, and the Ph.D. degree from the Technical University of Munich, Germany. His Ph.D. thesis was related to water resources management, climate change, and artificial intelligence in Mekong River Delta. He is currently a Postdoctoral Researcher with the Institute of Applied Sciences (HIAS), Ho Chi Minh City University of Technology (HUTECH). He has collaborated actively with researchers in several other disciplines of computer science, climatology, and computational data. He is in charge of the Research Group of Artificial Intelligence and Water Resources Engineering, as a Coordinator. He also collaborates with many colleagues from U.K., The Netherlands, Singapore, and USA. His research interests include hydrological, hydrodynamic modeling, downscaling and climate change, and artificial intelligence.

SON T. MAI is an Assistant Professor at Queen's University Belfast, U.K. Prior to this, he was at the University of Grenoble Alpes, France; Aarhus University, Denmark; and Ludwig—Maximilians University of Munich (LMU), Germany. His research interests include machine learning and data mining algorithms for large complex data.

ANH LE received the Ph.D. degree in computer science, University of Heidelberg, Germany, in 2014. He is currently the Dean of the Faculty of Information Technology, Ho Chi Minh City University of Transport, Vietnam. He has strong expertise in developing applications for intelligent transportation systems using data mining and machine learning. His research interests include spatio-temporal data mining, machine learning, and artificial intelligence.

VAN-HAU NGUYEN received the degree in applied informatics mathematics and the master's degree in information technology from the Hanoi University of Science and Technology, in 2003 and 2006, respectively, and the Ph.D. degree in computer science from the Artificial Intelligence Laboratory, Technische Universitäte Dresden, Germany. He is currently the Director of the AI Center, Faculty of Information and Technology, Hung Yen University of Technology and Education, Vietnam. He has published over 30 articles in international conferences and journals. His research interests include Boolean satisfiability problems (SAT), automated reasoning, machine learning, artificial intelligence, and deep learning.

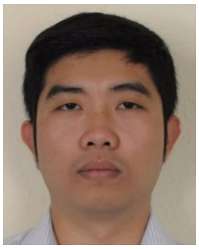


VAN-CHIEN NGUYEN is currently pursuing the bachelor's degree with the Hanoi University of Science and Technology. His research interests include AI technologies, machine learning, deep learning, reinforcement learning, and graph neural networks.

NGUYEN XUAN TINH received the Ph.D. degree in civil and environmental engineering from Tohoku University, Japan, in 2009. He is currently an Assistant Professor at Tohoku University. He has been awarded three prestigious prizes: the Best Paper Award at the Coastal Engineering Journal, in 2020; the JAMSTEC Nakanishi Award, in 2021; and the JSCE Best Paper Award, in 2021. His current research interests include river and coastal engineering, hydrodynamic modeling, ocean energy, natural disaster prevention and mitigation, climate changes, and water environmental engineering. Since 2009, he has published more than 30 journal articles in these topics. He is a member of the Japan Society of Civil Engineers (JSCE) and the International Association for Hydraulic Engineering and Research (IAHR).

HITOSHI TANAKA received the Ph.D. degree from Tohoku University, Japan, in 1984. After job experience in other universities, such as Utsunomiya University, Japan, and the Asian Institute of Technology, Bangkok, Thailand, he was promoted to Full-Professor at Tohoku University, in 1996. He was a Visiting Researcher at the Institute of Hydrodynamics and Hydraulic Engineering (ISVA), Technical University of Denmark, in 1996. His research interests include fluid mechanics, such as turbulent wave boundary layers, related sediment movement, and also resulting morpho-dynamics in costal and estuarine environments. His study sites are not confined in Japan, but covering various countries worldwide, such as Vietnam, Thailand, South Korea, Indonesia, Oman, and Bolivia. He served as the Chairman of the Asian and Pacific Division, International Association for Hydro-Environment Engineering and Research (IAHR-APD), from 2011 to 2014, and also as a Council Member of IAHR, from 2013 to 2017. From 2015 to 2017, he was a Vice-President of Japan Society of Civil Engineers (JSCE).

NGUYEN TRUNG VIET received the Ph.D. degree from Tohoku University, Japan, in 2007. He is currently a Professor at Thuyloi University, Vietnam. He has strong expertise in nearshore hydro-morphodynamics using both field measurements and numerical modeling, and remote sensing video techniques. He was the project leader of numerous MOST projects on Nha Trang, from 2013 to 2019, and Cua Dai Beach, from 2015 to 2018, in very close collaborations with IRD/EPOC/AFD-France, Japan, and U.K. He published more than 100 articles in journals and international conferences. He has been promoted to Vice President of Thuyloi University, since 2014. He is appointed as a Distinguished Member of IAHR-APD and an Executive Member of Asian and Pacific Coasts (APAC) Council. He has been a member of the State Council for Professorship, Vietnam, in the field of hydraulic engineering, since 2019.



LONG D. NGUYEN (Member, IEEE) was born in Dong Nai, Vietnam. He received the B.S. degree in electrical and electronics engineering and the M.S. degree in telecommunication engineering from the Ho Chi Minh City University of Technology (HCMUT), Vietnam, in 2013 and 2015, respectively, and the Ph.D. degree in electronics and electrical engineering from Queen's University Belfast (QUB), U.K., in 2018. He was a Research Fellow at QUB, for a part of Newton project, from 2018 to 2019. He is currently with Dong Nai University, Vietnam, as an Assistant Professor, and Duy Tan University as an Adjunct Assistant Professor. His research interests include convex optimization techniques for resource management in wireless communications, energy efficiency approaches (heterogeneous networks, relay networks, cell-free networks, and massive MIMO), and real-time embedded optimization for wireless networks and the Internet of Things (IoT).

TRUNG Q. DUONG (Senior Member, IEEE) is currently a Professor and the Chair of telecommunications at Queen's University Belfast, U.K., and the Research Chair of the Royal Academy of Engineering. His current research interests include wireless communications, signal processing, machine learning, and data analytics. He was awarded the Best Paper Award at the IEEE Vehicular Technology Conference (VTC-Spring), in 2013; the IEEE International Conference on Communications (ICC) 2014; the IEEE Global Communications Conference (GLOBECOM) 2016; the IEEE Digital Signal Processing Conference (DSP) 2017; and GLOBECOM 2019. He was a recipient of the prestigious Royal Academy of Engineering Research Fellowship, from 2015 to 2020, and has won the prestigious Newton Prize, in 2017. He currently serves as an Editor for the IEEE Transactions on Wireless Communications and an Executive Editor of IEEE Communications Letters.