Journal of Sea Research 198 (2024) 102472

Contents lists available at ScienceDirect

Journal of Sea Research

journal homepage: www.elsevier.com/locate/seares

Time series prediction of sea surface temperature based on BiLSTM model with attention mechanism

Nabila Zrira [a], Assia Kamal-Idrissi [b], Rahma Farssi [c], Haris Ahmad Khan [d,e,*]

[a] National Superior School of Mines Rabat, LISTD Laboratory, ADOS Team, Rabat, Morocco
[b] Mohammed VI Polytechnic University, Ai movement, Center of Artificial Intelligence, Rabat, Morocco
[c] National Superior School of Mines Rabat, Morocco
[d] Agricultural Biosystems Engineering, Wageningen University & Research, Wageningen, the Netherlands
[e] Data Science, Crop Protection Development, Syngenta, the Netherlands

ARTICLE INFO

Keywords:
Bidirectional Long Short-Term Memory (BiLSTM)
Attention
Sea Surface Temperature (SST)
Prediction
Marine data
Morocco

ABSTRACT

With the advancement of technology, ocean observation techniques have become increasingly prevalent in estimating marine variables such as Sea Surface Temperature (SST). This progress has led to a substantial surge in the volume of marine data. Presently, the abundance of available data presents a remarkable opportunity for training predictive models. The prediction of SST poses a challenge due to its temporal-dependent structure and multi-level seasonality. In this study, we propose a deep learning approach that combines the Bidirectional Long Short-Term Memory (BiLSTM) model with the attention mechanism to forecast SST. By leveraging the BiLSTM's ability to effectively capture long-term dependencies through both forward and backward LSTM processing, the attention mechanism accentuates salient features, thereby enhancing the model's evaluation accuracy.

To evaluate the effectiveness of the Attention-BiLSTM model in predicting SST, we conducted a case study in the Moroccan Sea, focusing on four distinct regions. We compared the performance of the Attention-BiLSTM model against alternative models such as LSTM, Attention-BiGRU, XGBoost, Random Forest (RF), Support Vector Regression (SVR), and Transformers in forecasting the SST time series.

The experimental results unequivocally demonstrate that the Attention-BiLSTM model achieves significantly superior prediction outcomes and is a good candidate for deployment in the field.

1. Introduction

Over the past two decades, satellite instruments and in situ observations, including marine stations, buoys, and voluntary observing ships, have provided valuable measurements of the physical characteristics of the ocean surface. These observations, despite their varying spatial and temporal sampling and measurement techniques, have become indispensable for understanding ocean dynamics. Notably, the use of altimetry data to study dynamic ocean topography has significantly advanced our comprehension of medium- to large-scale ocean variability (over 100–200 km) (Wang et al., 2022).

Sea Surface Temperature (SST) is a complex ocean parameter with significant implications for climate and marine ecosystems. Analyzing SST enables the management of marine ecosystems, as it provides insights into ocean conditions and climatic dynamics (Xuan et al., 2020). Accurate prediction of SST is of utmost importance, ranging from short-term to long-term forecasts. These predictions aid decision-making processes related to tasks such as fisheries resource distribution and marine environmental protection. Consequently, SST prediction plays a critical role in various domains, including marine fisheries, weather forecasting, marine animal tracking, and ship path planning.

Recent literature has witnessed an active research focus on SST prediction. This field involves generating future SST values in advance using predictive models with different time horizons (e.g., minutes, hours). However, SST data is often incomplete and structured as a time series, featuring complex or irrelevant coupling mechanisms and spatio-temporal relationships. These characteristics pose challenges to traditional data analysis approaches. To overcome these limitations, researchers have developed Machine Learning (ML) and Deep Learning (DL) models, which have demonstrated greater robustness and accuracy.

DL methods, such as Recurrent Neural Networks (RNNs), have been employed in previous studies to predict SST time series. Long Short-

* Corresponding author at: Agricultural Biosystems Engineering, Wageningen University & Research, Wageningen, the Netherlands.
E-mail address: haris.khan@wur.nl (H.A. Khan).

https://doi.org/10.1016/j.seares.2024.102472
Received 5 October 2023; Received in revised form 2 January 2024; Accepted 18 January 2024
Available online 24 January 2024
1385-1101/© 2024 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

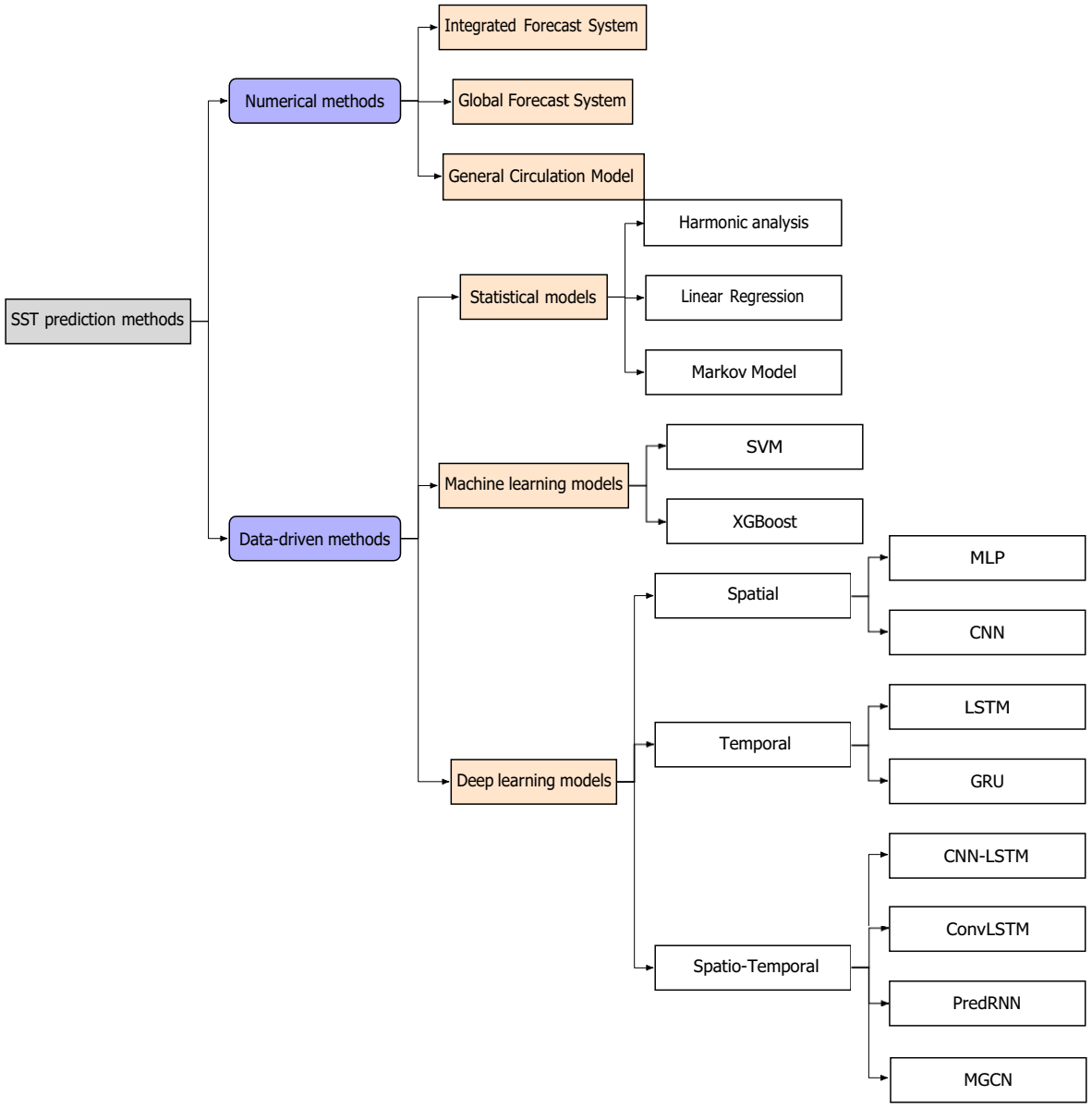

Fig. 1. Taxonomy of SST prediction models.

Term Memory (LSTM), a type of RNN, is particularly popular due to its ability to retain information over long periods of time. Numerous variations of LSTM have been proposed to enhance prediction performance by leveraging historical data to predict SST (Kun et al., 2021; Fei et al., 2022). The attention mechanism assigns weights to historical data, enhancing the utilization of historical temporal information. However, these versions of LSTM process sequences in one direction, limiting their ability to utilize future information.

To address this limitation, a Bidirectional Long Short-Term Memory neural network (BiLSTM) was developed, which consists of two LSTM models operating in both backward (future to past) and forward (past to future) directions. This architecture combines the advantages of sequential data processing and the long-term memory capabilities of both forward and backward LSTMs (Nie et al., 2021).

This paper proposes an attention-biLSTM model that effectively captures the relationships between historical and future SST values by employing a robust attention mechanism. The combination of attention and BiLSTM has demonstrated its effectiveness in similar time series prediction tasks in some previous works (Ahmed et al., 2022; Lee et al., 2022; Yang and Wang, 2022; Ma et al., 2020). Previous studies on marine data in Morocco have primarily focused on forecasting seasonal precipitation (Tuel and Eltahir, 2018). To the best of our knowledge, this is the first attempt to utilize the BiLSTM with an attention mechanism model for SST prediction in marine data, specifically in the context of Morocco. The main contributions of this paper are:

- The use of BiLSTM with optimal hyperparameters using K-fold cross-validation captures good temporal patterns in SST time series forecasting.
- The use of the attention mechanism allows BiLSTM to focus on relevant temporal patterns and enhances its ability to capture complex relationships within time series data. Potentially, it improves forecasting accuracy and provides valuable interpretability for SST time series data.
- A series of experiments in four Moroccan cities were conducted to evaluate the proposed model. Also, the attention-BiLSTM model was compared to other recurrent neural networks, traditional machine learning models, a gradient boosting model, and the recent deep learning model. In all experiments, the proposed model showed the highest prediction results.

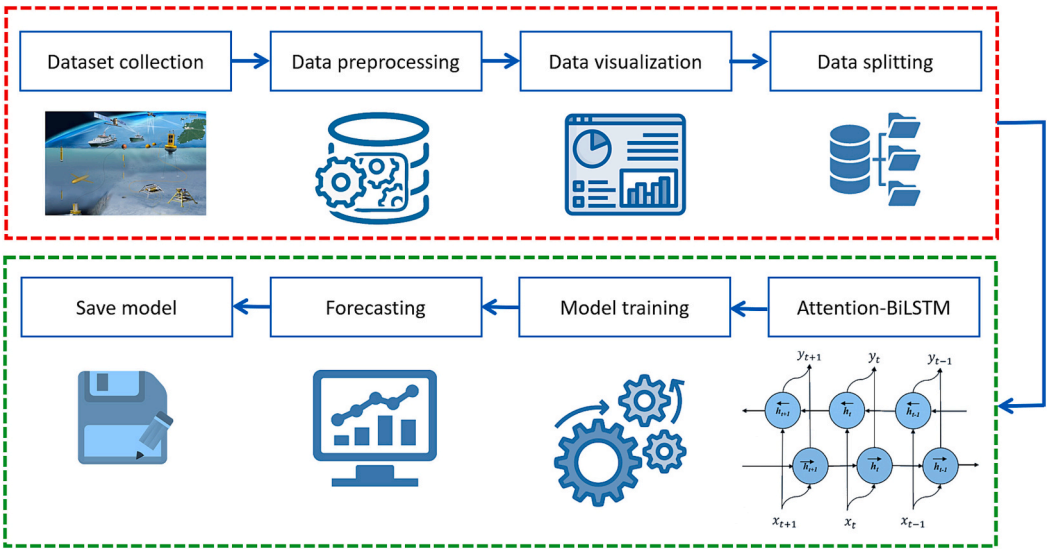

Fig. 2. Flowchart of our proposed approach.

The remainder of this paper is organized as follows. Section 2 provides a review of relevant literature on the time series prediction of SST. Section 3 introduces the proposed BiLSTM with an attention method for SST prediction. Section 4 describes the study area and presents the experimental results. Finally, Section 5 concludes the paper.

2. Related work

The prediction methods of SST can be roughly divided into two types: numerical methods and data-driven methods, where data-driven models can be further categorized into statistical models, machine learning models, and deep learning models. Data-driven methods encompass various statistical and artificial intelligence techniques, including ML methods (Ali et al., 2021). In contrast to numerical models, data-driven approaches require more data but less prior knowledge, resulting in improved prediction accuracy. Fig. 1 presents the taxonomy of SST prediction models.

2.1. Numerical methods

Real-time forecasts of SST are provided by many agencies around the world, namely the European Centre for Medium-Range Weather Forecasts (ECMWF) and the Command National Centers for Environmental Prediction (NCEP). These methods of SST prediction consist of modeling the physical environment parameters based on kinetic and thermodynamic equations (Xie et al., 2019). They are widely used in large seas. Many authors have used coupled General Circulation Models (GCMs) to predict SST over a basin. Krishnamurti et al. (2006) used thirteen state-of-the-art coupled global atmosphere-ocean models to predict seasonal global SST anomalies.

Besides, the Integrated Forecast System (IFS) is used by ECMWF and the Global Forecast System (GFS) by NCEP, respectively. The IFS forecasts the SST of the following 10–15 days while the NCEP forecasts the SST of the following 380 h (Aparna et al., 2018).

2.2. Data-driven methods

Data-driven prediction methods use statistical methods to predict SST. This kind of method requires a large amount of data but less knowledge of the oceans and atmosphere due to its capability to learn patterns from historical data automatically and further use the learned patterns to predict future values of SST. This category of methods includes artificial intelligence, namely machine learning and deep learning models. They outperform highly complex computations with excellent time series prediction results (Shi et al., 2022).

2.2.1. Statistical methods

Statistical methods learn from historical data to predict SST. They range from simple regression, such as linear regression, to complex methods like the Markov model (Xue and Leetmaa, 2000). Autoregressive Integrated Moving Average (ARIMA), which is a Regression model, has been widely used for SST predictions due to its simplicity, adaptability, and the Box and Jenkins methodology (de Mattos et al., 2022). The latter is used to provide a well-established design process for linear patterns.

2.2.2. Machine Learning methods

With the continuous development of technology, ML models have been efficiently used as an alternative to statistical methods for different tasks due to their nonlinearity, performance, and flexibility. They have shown an important capacity and great potential to predict the parameters of the ocean and weather. Among these models are Support Vector Machines (SVM) and XGBoost. The latter is an option to describe SST's characteristics, which are complex temporal-dependence structure and multi-level seasonality (Wolff et al., 2020). Meanwhile, SVM constructs a set in a high-dimensional space that can be used in regression task (Hou et al., 2022). For example, Lins et al. (2013) proposed an SVM combined with multi-linear regression (MR) to predict SST over the Northeastern Brazilian Coast and the tropical Atlantic.

Besides, the Support Vector Regression (SVR) model achieves a satisfactory result in predicting SST. Basically, these ML models outperform the numerical models in terms of accuracy and simplicity (Lee et al., 2016; Rehana, 2019). However, these kinds of models are not able to properly model nonlinear patterns in temporal phenomena.

2.2.3. Deep Learning methods

As ML models, Deep Learning models have recently experienced rapid development due to their ability to model complex relationships by extracting features from hidden data. They are widely used in many applications, such as time series prediction tasks. Choi et al. (2022) proposed a new hybrid method to enhance the accuracy of a real-time ocean forecasting system by combining a deep Generative Inpainting Network (GIN) and a numerical ocean model. The most exciting studies specifically focused on temperature-related tasks using CNNs have been those that employed satellite data. Since the prediction of SST is basically a time series problem, the use of the LSTM architecture is common

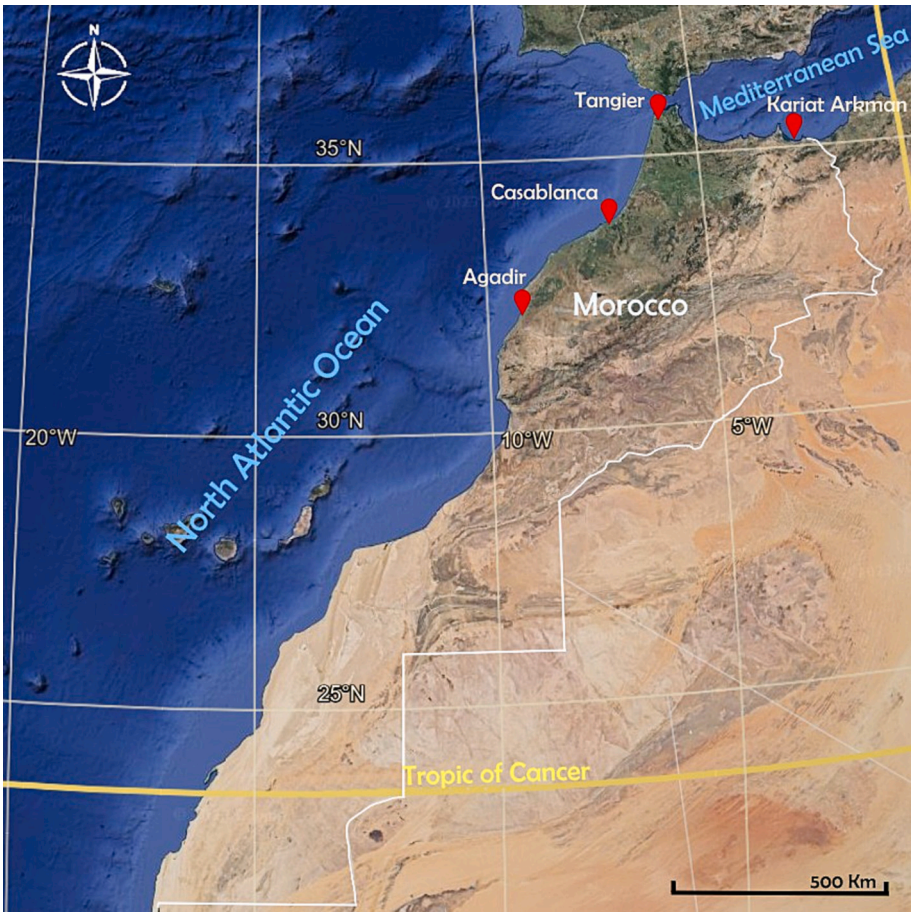

Fig. 3. The selected Moroccan cities.

for this task. Xiao et al. (2019) used a combination of CNN and LSTM, which is a spatio-temporal deep learning model, to forecast SST. The proposed model used satellite data, rain gauge data, and thermal infrared images. To extract spatial features, the CNN model was used for this objective, meanwhile, the LSTM model aims to handle the time dependencies of the provided data. Results show that the CNN-LSTM model outperforms comparative models, such as CNN, LSTM, and MLP. They had relatively accurate prediction results for the short-term and mid-term daily forecast of SST (Sit et al., 2020). The MLP was exploited by Wolff et al. (2020) to predict the short and long-term evolution of the SST. The algorithm was able to forecast SST with hindcast input and atmospheric data comparable to a state-of-the-art physics-based model simulation from the European Center for Medium Weather Forecasting.

Zheng et al. Zheng et al. (2020) proposed a model combining Deep Neural Network (DNN) and bias correction maps. DNN is composed of four stacked composite layers. Each layer processes SST maps at different resolutions to analyze systems at multiple scales simultaneously. The forecast results were validated by satellite observations in the eastern equatorial Pacific Ocean. Zhang et al. Zhang et al. (2021) designed a new Memory Graph Convolutional Network (MGCN) that captures spatio-temporal changes of SST. MGCN consists of two layers: the memory layer and the graph layer. The first layer encodes temporal changes using gate linear units and convolution units. Whereas, the second layer captures spatial changes in the frequency domain using Laplacian. More recently, Qiao et al. Qiao et al. (2023) suggested an ensemble learning method that combines a spatio-temporal deep learning network with the attention mechanism. The authors used the XGBoost algorithm to extract seasonal periodic features from SST data. Then, they extracted spatio-temporal features using PredRNN (Predictive Recurrent Neural Network). Finally, they added an attention mechanism to extract the important patterns in historical SST data and further improve the prediction accuracy. All these studies have shown that deep learning techniques have the advantages of strong nonlinear mapping and multidimensional information processing.

3. Methodology

The suggested pipeline for predicting future SST values is shown in Fig. 2. The approach consists of two major phases: data preparation and network training. First, we select the study region in which the SST values will be predicted. The data represents a time series that may encounter several problems, such as missing values (or timestamps), and outliers. For that, a preprocessing step is required to handle these problems. Moreover, we normalize time series data to improve the training stability as well as the performance of the Attention-BiLSTM network. Then, the processed time series data is split into training and testing sets. Second, we establish the Attention-BiLSTM model to perform accurate time series prediction. Finally, the best model with the optimized hyperparameters is saved for further use.

3.1. Study area and data

Morocco is the northwesternmost country in Africa. Its sea surface spans the Mediterranean Sea and the Atlantic Ocean on the north and the west, respectively. In this work, SST value prediction was carried out in four different Moroccan cities: Tangier, Agadir, Casablanca, and Kariat Arkmane. As shown in Fig. 3, the choice of these cities is based on their location relative to the sea. Agadir is located on the shore of the Atlantic Ocean near the foot of the Atlas Mountains. Casablanca is the

Table 1
GPS coordinates of the study area.

Region	Coordinates
Tangier	36.25026152534847, −5.977397294819712
Agadir	31.14608750019331, −9.66901533235119
Casablanca	33.97723556735074, −7.843945740762594
Kariat Arkmane	35.12039529923372, −2.734379316487201

Table 2
Summary statistics of SST data.

Region	Mean	Warmest	Coldest	STD
Tangier	18.360108	24.19	14.62	2.266212
Agadir	18.349655	22.77	14.54	1.692106
Casablanca	19.106891	24.17	14.79	2.373244
Kariat Arkmane	19.549607	27.63	13.76	3.709784

largest city in Morocco. It is located on the Atlantic coast of the Chaouia Plain, in the central-western part of Morocco. Kariat Arekmane is a coastal town in Morocco, in the Oriental region. It is located in the Nador province, 25 km from the city of Nador. Kariat Arekmane is served by the Mediterranean Ring Road, which crosses it. Tangier is a city in northwestern Morocco. The Atlantic Ocean meets the Mediterranean Sea on the coast at the western entrance to the Strait of Gibraltar. Table 1 depicts the GPS coordinates of each selected city in the study area.

The measurements of the SST are provided by the daily satellite readings provided by the National Oceanic and Atmospheric Administration (NOAA) from January 01, 2012, to August 24, 2022. Table 2 summarizes statistics related to SST data, including measures of the mean, standard deviation, and specific extreme values such as minimum and maximum temperature. The choice of these four cities also depends on statistics, which are similar in most cities. In deep learning, using data with similar statistics can help in building models that generalize well. Models tend to perform better when applied to new, unseen data that follows the same distribution. Moreover, similar statistics can be beneficial when splitting data into training and testing sets. It ensures that both sets have comparable statistical properties, which helps in creating reliable models and assessing their performance accurately.

3.2. Data preprocessing

Predictive models for time series require clean and complete data without outliers or missing values. Therefore, it is crucial to apply appropriate preprocessing techniques before utilizing these models. In this study, we employed standard preprocessing methods to enhance the quality of the data, including conventional imputation for handling missing values.

Furthermore, it is essential to normalize the time series data due to its varying measurement scales. Deep learning methods aim to map input data to output across all samples in the training set. The model initializes the weights randomly and updates them using optimization techniques like Stochastic Gradient Descent (SGD) or Root Mean Square Propagation (RMSProp). The initial weights and the error calculated between the predicted and actual values emphasize the significance of scaling both the input and output. Failure to scale the time series data appropriately in prediction problems can lead to an unstable learning process.

To handle this problem, we perform a min-max normalization technique to transform time series data in the range $(0, 1)$. Equation1 depicts the feature scaling formula:

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

Where x represents the origin value, x_{min} and x_{max} are the lowest and highest values, respectively.

3.3. Data splitting

The time series data used in this study spans a period of 11 years, specifically from January 1, 2012, to August 24, 2022, encompassing a total of 3889 days. To divide the dataset into training and testing sets, the data from January 1, 2012, to December 31, 2021 (a total of 3653 days) is allocated for training purposes. The remaining data, consisting of 236 days, is reserved for testing the model's performance.

For a visual representation of the data distribution across the study area, please refer to Fig. 4. This figure illustrates the training and testing data for each region within the study area.

3.4. Bidirectional Long Short-Term Memory (BiLSTM)

Long Short-Term Memory (LSTM) is a specialized type of Recurrent Neural Network (RNN) that incorporates Long Short-Term Memory units. These units were introduced by Hochreiter and Schmidhuber, German researchers, as a solution to the vanishing gradient problem (Hochreiter and Schmidhuber, 1997). LSTMs have proven to be highly effective across a wide range of problem domains and are now widely utilized in various applications.

LSTMs consist of memory blocks, which contain memory cells with self-connections. These cells retain the temporal state of the network. Additionally, LSTMs incorporate multiplicative units called gates, which enable the storage, retrieval, and modification of information within the cells. Each cell autonomously determines which information to store and controls the opening and closing of gates for reading, writing, and resetting. The gates operate in a binary manner, with two states: open and closed. However, these gates are implemented using element-wise multiplication by the sigmoid function, which restricts their values to the range of 0 to 1. This analog implementation enables differentiability and facilitates the backpropagation algorithm. The gates process the signals they receive, selectively allowing or blocking information based on its relevance and importance, which is determined by their respective weights. These weights are adjusted during the recurrent network learning process, modulating the input and hidden states of the LSTM.

First, the LSTM unit relies on deciding what information seems important to keep. The decision is made by the forget layer, which looks at x_t and h_{t-1}, then outputs a value between 0 and 1 in the cell state c_{t-1} using a sigmoid function.

$$f_t = \sigma\left(w_f \cdot x_t + u_f \cdot h_{t-1} + b_f\right) \tag{2}$$

Second, the LSTM unit selects the new information that should be stored in the cell state. The input gate layer i_t decides which values will be updated using the sigmoid function. Then, a *tanh* function creates a vector of new candidate values $\widetilde{c}$ that could be added to the state.

$$i_t = \sigma(w_i \cdot x_t + u_i \cdot h_{t-1} + b_i) \tag{3}$$

$$\widetilde{c}_t = tanh(w_c \cdot x_t + u_c \cdot h_{t-1} + b_c) \tag{4}$$

Third, the old state c_{t-1} is updated into the new cell state c_t.

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \widetilde{c}_t \tag{5}$$

Finally, the output gate o_t decides the amount of memory content to yield to the next hidden state.

$$o_t = \sigma(w_o \cdot x_t + u_o \cdot h_{t-1} + b_o) \tag{6}$$

$$h_t = o_t \cdot tanh(c_t) \tag{7}$$

Where $(\cdot)$ is the inner products, $w_{()}$, and $u_{()}$ are the weights, and $b_{()}$ is the bias.

As shown in Fig. 5, BiLSTM represents an amelioration of the LSTM model that contains two separate LSTM networks: backward and forward (Graves and Schmidhuber, 2005; Graves et al., 2005). The forward hidden state is calculated by Eq. (8):

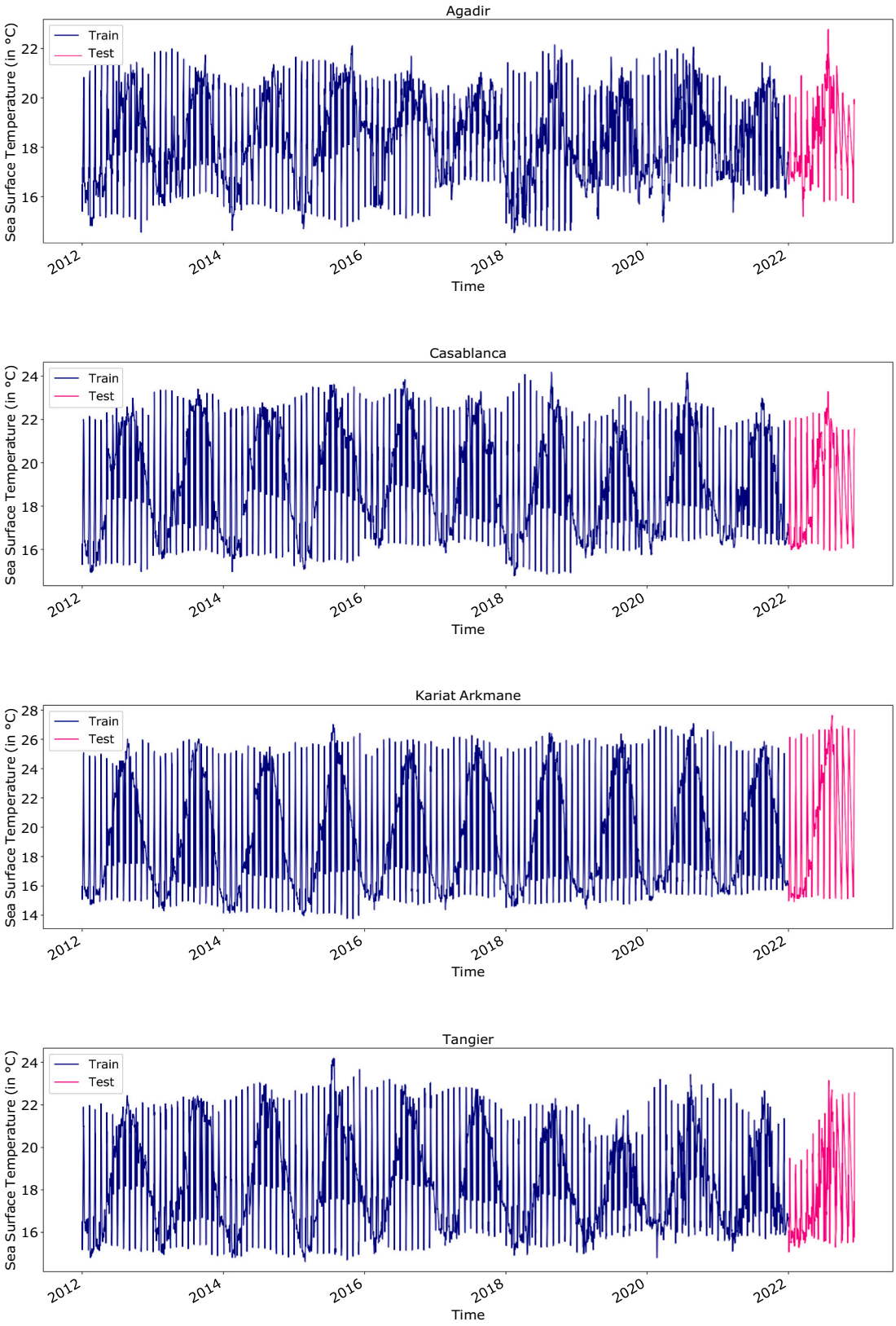

Fig. 4. Training and testing data.

$$\overleftarrow{h} = o_t \cdot tanh(c_t) \tag{8}$$

Where: o_t represents the output gate and c_t represents the cell state. Similarly, the backward hidden state $\overleftarrow{h}$ is calculated to the forward layer. Both layers are concatenated and fed forward to the next layer. BiLSTM ensures the integration of the past and future information for each point. In Eq. (9) element-wise sum is used to concatenate the forward and backward pass outputs.

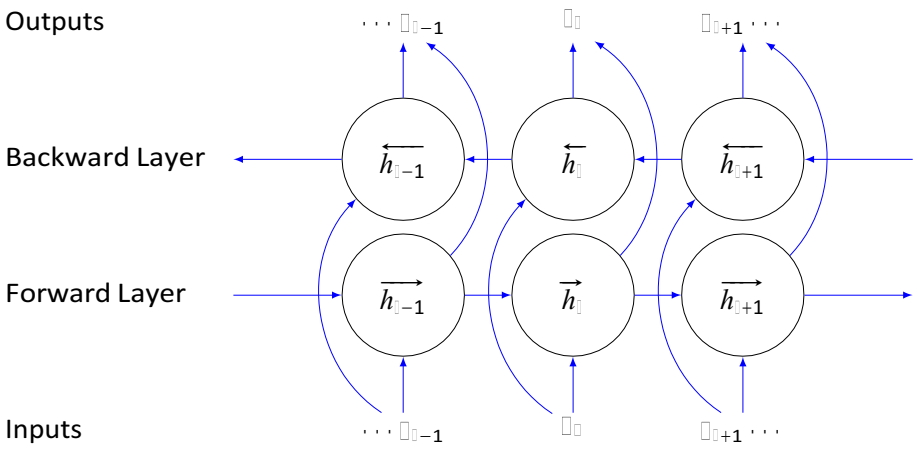

Fig. 5. BiLSTM architecture.

Table 3
Explanation of BiLSTM parameters.

Parameter	Explanation
x_t	The input vector at time t (i.e., SST time series)
f_t	The output of the forget gate
i_t	The output of the input gate
$\widetilde{c}_t$	The vector of candidate values that should be added to the cell state
c_t	The new cell state
o_t	The output of the output gate
h_t	The hidden state at time t
y_t	The final predicted output at time t

$$h_t = \overleftarrow{h_t} \oplus \overrightarrow{h_t} \tag{9}$$

Finally, an activation function is applied to the hidden state h_t in order to generate the final output y_t. Table 3 gives more explanation about all parameters of the BiLSTM equations.

3.5. Attention mechanism

The attention mechanism of humans leads us to selectively focus on the important information required and ignore other visible information that seems irrelevant (Niu et al., 2021). Nowadays, attention is extensively applied to computer vision (Wang and Tax, 2016), natural language processing (Galassi et al., 2020), machine translation (Bahdanau et al., 2014; Luong et al., 2015), time series prediction (Noor et al., 2022), and other fields (Yang et al., 2016; Wang et al., 2016).

The attention mechanism helps models to provide an ability to focus only on crucial features from all the data. Deep Learning adopted the attention mechanism in time series models to improve the accuracy of predicting.

The proposed approach adds an attention layer after three BiLSTM layers. The output of the last BiLSTM layer represents the input of the attention layer that will assign more weights to SST features that seem important in predicting. At this stage, the approach computes the probability according to the weight distribution. Then, it updates and optimizes weight parameters at each iteration during training. The attention mechanism can be expressed by the following formulas (10,11,12):

$$A_t = v \quad tanh(wh_t + b) \tag{10}$$

$$\alpha_t = \frac{exp(A_t)}{\sum_{j=1}^{i} A_j} \tag{11}$$

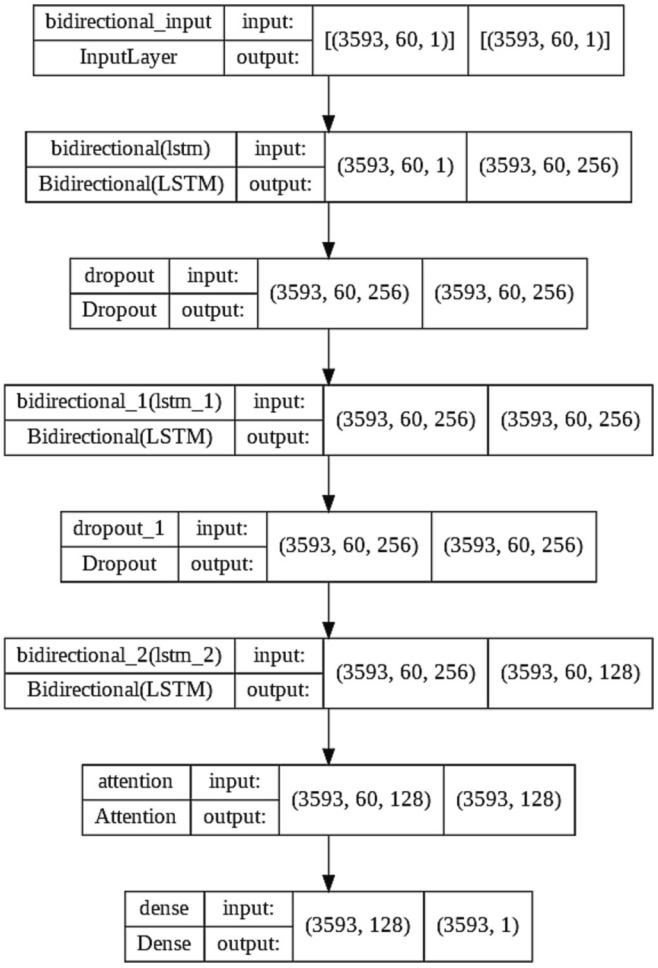

Fig. 6. Attention-BiLSTM architecture.

$$O_t = \sum_{t=1}^{i} \alpha_t h_t \tag{12}$$

Where:

- A_t: represents the attention probability distribution;
- v and w: represent the weight coefficients;
- b: represents the bias coefficient;

- O_t: represents the attention layer output at time t.

To select the number of layers and neurons in our model, we performed a k-fold cross-validation (5 fold), which plays a crucial role in assessing model performance, tuning hyperparameters, and evaluating generalization ability. This technique involves systematically exploring various architectures and hyperparameters to find the best combination. The best architecture of the Attention-BiLSTM model is depicted in Fig. 6, which comprises several layers. It begins with an input layer, followed by three bidirectional LSTM layers, two dropout layers, an attention layer, and a dense layer. The input layer has a shape of $(3593, 60, 1)$, where 3593 represents the training data and 60 represents the number of time steps. Determining the optimal number of time steps in neural networks, particularly for LSTM models, is not straightforward. Therefore, we conducted several experiments using RMSE and R^2 metrics to identify the most suitable value.

Each bidirectional LSTM layer consists of 128, 128, and 64 units, respectively. Dropout layers are employed to mitigate overfitting. In our experiments, LSTM units were randomly deactivated during training, with a dropout rate of 20%.

The weight kernel was initialized using Glorot uniform (Glorot and Bengio, 2010) to prevent the activation outputs of the layers from experiencing gradient explosion or vanishing during the forward pass.

4. Experimental results

In this section, we deploy qualitative and quantitative results to demonstrate the efficiency of the Attention-BiLSTM on SST value predicting. First, we introduce a brief description of evaluation metrics as well as the experimental parameter settings, and then we compare the performance of Attention-BiLSTM with the standard LSTM, Transformers, and other forecasting methods.

4.1. Evaluation metrics

To analyze and select the best-performing models, the most common evaluation metrics used in time series predicting are adopted. Given x the actual observation, y the predicted observation, and N the total number of testing set, the evaluation criteria are computed based on the following eqs.

A Mean Absolute Error (MAE) represents the average of absolute errors of prediction for a set of observations and predictions.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|x_i - y_i| \tag{13}$$

A Mean Absolute Percentage Error (MAPE) represents the mean absolute percentage error.

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{x_i - y_i}{x_i}\right| \tag{14}$$

A Root Mean Squared Error (RMSE) is calculated as the root mean square error. The RMSE is very used for numerical predictions.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|x_i - y_i|^2} \tag{15}$$

R-squared (R^2) (also called the coefficient of determination) is more informative than MAE, MAPE, and RMSE metrics. R^2 is the proportion of the variance for a dependent variable that's explained by an independent variable.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}|x_i - y_i|^2}{\sum_{i=1}^{N}|x_i - \overline{x_i}|^2} \tag{16}$$

Table 4
Hyperparameters maintained during deep learning training process.

Hyperparameter	Value
Input shape	(3593, 60, 1)
Units	128–128-64
Batch size	32
Learning rate	10^{-4}
Optimizer	Adam
Epochs	150
Loss function	Mean Squared Error
Dropout	0.2
Kernel initializer	Glorot uniform

Table 5
The SST prediction for different time steps in Kariat Arkmane.

Time steps (in days)	MAE	MAPE	RMSE	R^2
1	0.225811	0.011401	0.290551	0.995060
7	0.220457	0.010779	0.301437	0.994683
15	0.246256	0.011746	0.342768	0.993125
30	0.215287	0.010426	0.296430	0.994858
45	0.204129	0.009934	0.280976	0.995380
60	0.204140	0.010046	**0.280905**	**0.995383**
75	0.217545	0.010923	0.284448	0.995265
90	0.206601	0.010186	0.283890	0.995284

Where $\overline{x_i}$ represents the mean of x observations.

4.2. Experiment settings

The network development for this study was conducted in Google Colab, utilizing TensorFlow, Keras, and the Attention frameworks. The training and testing phases were executed on GPUs T4 and P100 within the same Python 3.5 programming environment.

Attention-BiLSTM is trained for approximately 150 epochs, utilizing an early stopping technique to prevent overfitting. If the loss function (Mean Squared Error) does not improve after 10 epochs, the learning rate is reduced. The optimization of the model is conducted using the Adam technique Kingma and Ba (2014) with a mini-batch size of 32. The initial learning rate is set to 10^{-4}.

For deep learning methods, i.e., the Attention-BiLSTM, the Attention-BiGRU, and Transformers, the hyperparameters used during training are presented in Table 4. These hyperparameters include the number of LSTM/GRU units, the dropout rate, the learning rate, and the batch size. These hyperparameters were selected based on experimentation and fine-tuning to achieve optimal performance for the Attention-BiLSTM and Attention-BiGRU models.

In recurrent neural networks, the time step refers to the number of time steps or past observations used to make predictions for the next time step in a sequence. However, it represents an important hyperparameter that might depend on the patterns and relationships within the data. For that reason, we conducted several experiments to select the optimal time step. Table 5 presents the results, showing that the highest R^2 value and the lowest RMSE value were obtained when using 60-time steps. Two months can capture changes related to seasonal variations in SST. Observing SST over this period allows for tracking how temperatures change as the seasons transition, which can be crucial for understanding SST patterns. Hence, we selected 60 as the number of time steps for this study.

The experimental results are presented and discussed to showcase the effectiveness of the proposed Attention-BiLSTM method. The time series data used in these experiments spans a total of 3889 days, starting from January 1, 2012, and ending on August 24, 2022. The training process utilizes data from January 1, 2012, to December 31, 2021, encompassing 3653 days. The remaining 236 days, from January 1,

Table 6
Ablation study of the Attention-BiLSTM model in Kariat Arkmane.

	Number of BiLSTM layers	Number of dropout layers
Model A	3	None
Model B	2	1
Model C	1	1
Model D	3	2

Table 7
The ablation study of our proposed network in Kariat Arkmane.

Model	MAE	MAPE	RMSE	R^2
Model A	0.311330	0.014951	0.433588	0.988999
Model B	0.233684	0.011443	0.324379	0.993843
Model C	0.324140	0.015948	0.453528	0.987964
Model D	**0.204140**	**0.010046**	**0.280905**	**0.995383**

Table 8
Parameters maintained during RF, SVR, and XGBoost training process.

Model	Parameter	Value
XGBoost	max_depth	3
	learning_rate	0.1
	n_estimators	100
RF	n_estimators	100
	criterion	absolute_error
SVR	degree	1
	kernel	linear

2022, to August 24, 2022, are reserved as the testing set for SST prediction.

All experiments are conducted on the testing set, using identical experimental settings. This ensures a fair comparison between different models and methods. By using the same testing set and experimental conditions, the results can be directly compared to evaluate the performance of the Attention-BiLSTM method.

To verify the performance of our proposed approach, we accomplish different scenarios:

- Model A: it represents the Attention-BiLSTM model with three bidirectional LSTM layers without using dropout layers;
- Model B: it represents the Attention-BiLSTM model with two bidirectional LSTM layers and one dropout layer;
- Model C: it represents the Attention-BiLSTM model with one bidirectional LSTM layer and one dropout layer;
- Model D: it represents the Attention-BiLSTM model with three bidirectional LSTM layers and two dropout layers.

Table 6 depicts an ablation study that involves systematically removing or disabling specific layers within our neural network to analyze their individual impact on the model's performance. The goal is to understand the significance of each layer and its contribution to the overall model performance.

In Table 7, we remark that dropout layers increase the achievement of our model by improving the generalization performance of Attention-BiLSTM. Also, the use of three bidirectional LSTM layers produced a stable model that provided good results.

4.3. Results and discussion

To assess the effectiveness of the proposed Attention-BiLSTM method, a benchmark comparison is conducted against six other methods. The benchmark includes two traditional machine learning models: Support Vector Regression (SVR) and Random Forest (RF). Additionally, a gradient boosting method called Extreme Gradient Boosting (XGBoost) is included. Three deep learning methods specifically designed for time series data, Long Short-Term Memory (LSTM), Attention-BiGRU (Bi-Gated Recurrent Unit), and Transformers are also part of the benchmark.

In the initial experiments, five evaluation metrics are used to compare the performance of these methods: Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), R^2, and execution time (in seconds) for both training and inference stages.

Furthermore, Table 8 provides a summary of the best parameters used during the training of the RF, XGBoost, and SVR models. These parameters are determined through experimentation and tuning to achieve optimal performance for each method.

Tables 9, 12, 11, and 10 provide insights into the performance of different models across various regions in Morocco. It is observed that the RF model generally performs the worst, exhibiting higher error values and a lower determination coefficient. On the other hand, the Attention-BiLSTM consistently achieves the highest R^2 in all Moroccan regions, indicating a good fit for the SST observations. Notably, the Kariat Arkmane region stands out with an R^2 of 0.995383, indicating highly accurate sea temperature predictions.

While training deep learning models can be computationally intensive and time-consuming, it is crucial to note that once a model is trained and the results are significant, it can be saved and used for inference. In terms of inference time, as shown in Tables 9, 10, 12, and 11, the Attention-BiLSTM model outperforms others in terms of accuracy, despite requiring a longer training process. On average, the Attention-BiLSTM model can forecast 230 days in less than 3 s, demonstrating its efficiency in making predictions in a timely manner.

Figs. 7, 10, 9, and 8 visually represent the predicted SST values for the Agadir, Casablanca, Kariat Arkmane, and Tangier regions, respectively. These plots provide further evidence of the quality of the predicted SST values by the Attention-BiLSTM model, as they closely align with the actual SST values.

The proposed Attention-BiLSTM model has demonstrated remarkable predictive performance across the testing set of four different Moroccan regions. The use of the attention mechanism in the model has contributed to improved predictions, as it allows the model to focus on important features. Compared to other models, the Attention-BiLSTM model proves to be a highly effective solution for predicting SST values.

Overall, the experimental results, visualizations, and comparative analysis indicate that the proposed Attention-BiLSTM method accurately predicts SST values in various Moroccan regions. The inclusion of

Table 9
The predicting performance on Agadir.

Model	MAE	MAPE	RMSE	R^2	Training	Inference
XGBoost	0.317217	0.017061	0.433130	0.927105	0.05	0.0010
RF	0.337023	0.018147	0.450764	0.921049	15.78	0.0107
SVR	0.318378	0.017125	0.425294	0.929719	2.543	0.0213
LSTM	0.312283	0.016706	0.442082	0.922751	437.38	0.3111
BiLSTM	0.306279	0.016511	0.419128	0.930565	446.21	1.4500
Attention-BiGRU	0.313244	0.016939	0.415326	0.931819	391.29	1.7513
Transformers	0.325090	0.017672	0.426211	0.928198	347.63	2.0101
Attention-BiLSTM	**0.300033**	**0.016150**	**0.413448**	**0.932434**	503.00	2.7600

Table 10
The Predicting performance on Tangier.

Model	MAE	MAPE	RMSE	R^2	Training	Inference
XGBoost	0.310400	0.017155	0.444201	0.954140	0.11	0.01
RF	0.315441	0.017460	0.454354	0.952020	14.09	0.01
SVR	0.298017	0.016496	0.424580	0.958102	1.40	0.01
LSTM	0.300582	0.016716	0.431939	0.955702	497.14	1.48
BiLSTM	0.298235	0.016482	0.420803	0.957957	487.34	1.45
Attention-BiGRU	0.299526	0.016551	0.424332	0.957248	357.57	1.75
Transformers	0.338829	0.018947	0.464905	0.948682	387.79	1.47
Attention-BiLSTM	**0.309356**	**0.017203**	**0.422227**	**0.957672**	453.72	2.03

Table 11
The Predicting Performance on Kariat Arkmane.

Model	MAE	MAPE	RMSE	R^2	Training	Inference
XGBoost	0.217379	0.010608	0.301351	0.994718	0.24	0.02
RF	0.225040	0.011079	0.302400	0.994681	23.83	0.01
SVR	0.212706	0.010399	0.289826	0.995114	3.18	0.01
LSTM	0.225529	0.011333	0.292770	0.994984	434.88	1.42
BiLSTM	0.212948	0.010609	0.284854	0.995252	392.92	1.21
Attention-BiGRU	0.206316	0.010158	0.282809	0.995320	393.35	1.83
Transformers	0.205884	0.010088	0.284455	0.995265	389.44	2.16
Attention-BiLSTM	**0.204140**	**0.010046**	**0.280905**	**0.995383**	453.7	2.11

Table 12
The Predicting performance on Casablanca.

Model	MAE	MAPE	RMSE	R^2	Training	Inference
XGBoost	0.170690	0.008851	0.243260	0.988788	0.05	0.001
RF	0.176252	0.009135	0.250328	0.988126	14.29	0.01
SVR	0.165814	0.008582	0.238048	0.989263	1.44	0.01
LSTM	0.203848	0.010782	0.263113	0.986738	446.78	1.83
BiLSTM	0.167853	0.008736	0.238249	0.989126	480.56	0.38
Attention-BiGRU	0.180324	0.009448	0.244791	0.988521	401.15	2.70
Transformers	0.172817	0.008909	0.244732	0.988526	342.30	1.93
Attention-BiLSTM	**0.168371**	**0.008775**	**0.237012**	**0.989239**	452.57	2.77

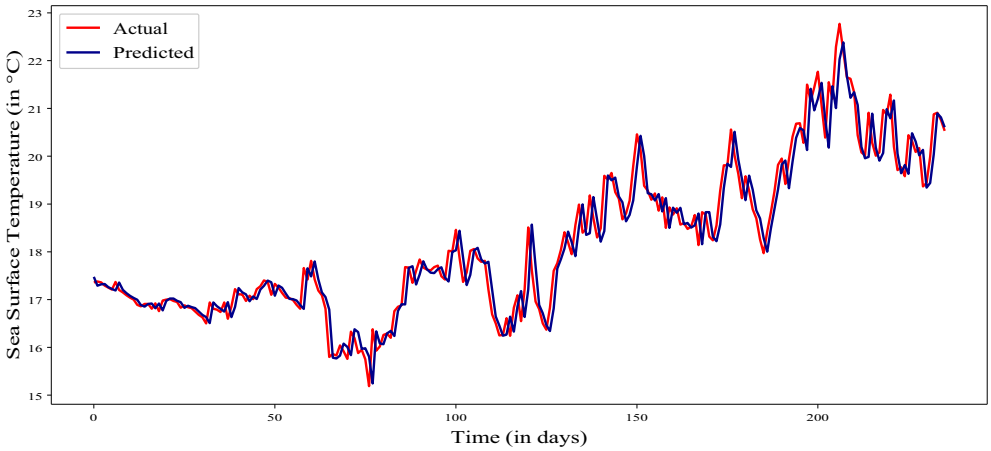

Fig. 7. Predicting results of Attention-BiLSTM on Agadir.

the attention mechanism enhances the model's performance, making it a superior choice compared to other models considered in the experiments.

Nevertheless, in this work, we encounter three limitations: the lack of availability of more data, the hyperparameter setting, and climate change. Dealing with limited data in time series analysis can indeed present significant challenges. The nature of time series data, often characterized by its sequential and temporal dependencies, can make the analysis complex, especially when there are few observations. Moreover, the hyperparameter setting is an essential part of the deep learning workflow, influencing a model's generalization, performance, and computational efficiency. It often requires a balance between the exploration of various settings and the computational resources available for experimentation. In these experiments, we used only four datasets from four Moroccan cities. However, the hyperparameter setting will be very complex when we extend our work to all seaside towns. Finally, climate change is a major and complex problem that can affect the prediction model. It refers to long-term changes in weather

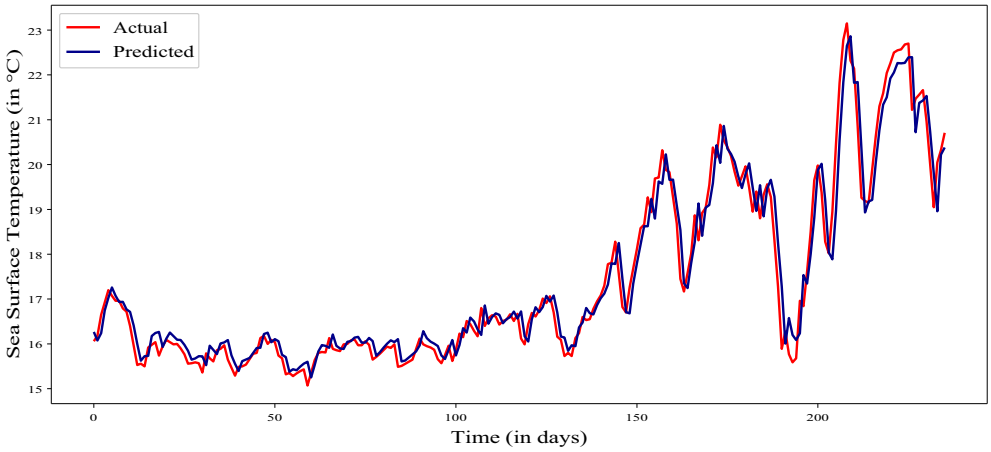

Fig. 8. Predicting results of Attention-BiLSTM on Tangier.

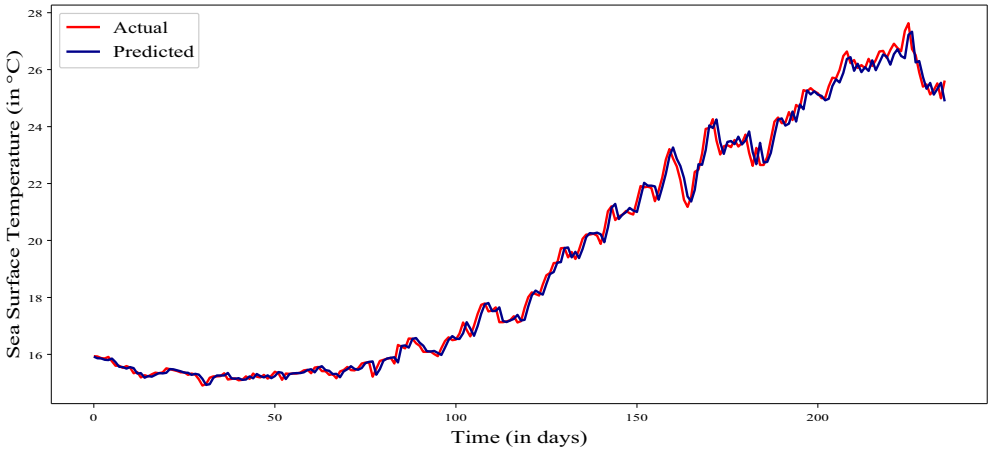

Fig. 9. Predicting results of Attention-BiLSTM on Kariat Arkmane.

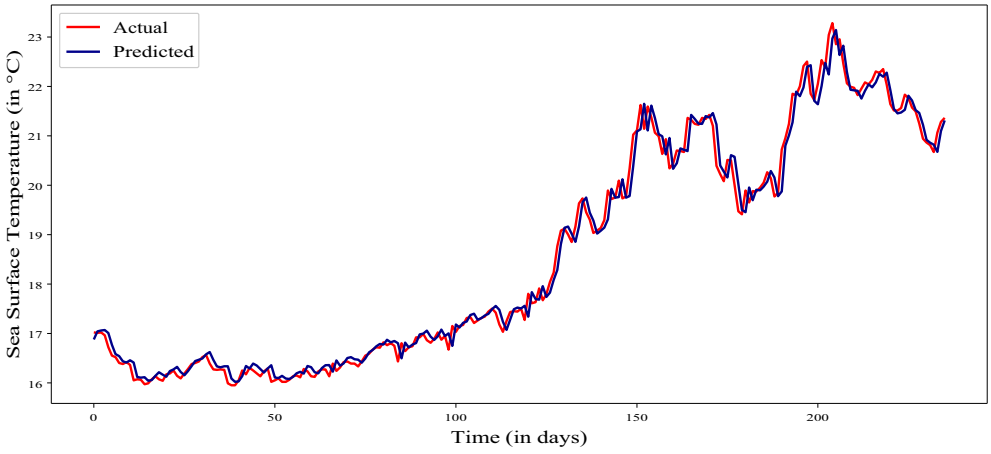

Fig. 10. Predicting results of Attention-BiLSTM on Casablanca.

patterns on a global or regional level. These changes include variations in average temperatures, removals, extreme weather events, and climate patterns.

5. Conclusion

In this paper, we propose a novel approach for sea surface temperature (SST) prediction utilizing deep learning techniques. Our method incorporates the Bidirectional Long Short-Term Memory (BiLSTM) deep recurrent neural network model along with the attention mechanism. The attention mechanism plays a crucial role in enhancing prediction accuracy by capturing relationships between historical and future SST values. Through extensive experimentation, we demonstrate that our proposed model outperforms existing SST prediction approaches. However, our method encounters some shortcomings. Indeed, our study relied only on four data sets provided by four Moroccan cities, but its

extension to all coastal cities considerably increases the complexity of tuning hyperparameters in deep learning models. Additionally, climate change emerges as a complex and influential factor impacting predictive models due to changes in global or regional weather patterns. The prediction of our model is very close to reality because it is currently learned on stable data. One of the most obvious impacts of climate change is increased SST. The oceans will absorb much of the excess heat trapped by greenhouse gases, causing SST to increase over time. In this case, our model will not be able to predict future SST values well.

In future work, we plan to explore the application of the neural prophet model to Moroccan SST data as well as other marine datasets such as Pirata (the Prediction and Research Moored Array in the Atlantic). By testing the neural prophet model, we aim to further enhance our understanding and predictive capabilities in the domain of SST forecasting. This will contribute to the advancement of marine research and enable better predictions for various regions and datasets.

CRediT authorship contribution statement

Nabila Zrira: Writing – review & editing, Writing – original draft, Methodology, Investigation, Data curation, Conceptualization. **Assia Kamal-Idrissi:** Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Rahma Farssi:** Writing – review & editing, Writing – original draft, Conceptualization. **Haris Ahmad Khan:** Writing – review & editing, Writing – original draft, Validation, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

The authors would like to thank the seatemperature website for providing Moroccan marine environment observation data https://seatemperature.info.

References

Ahmed, Dozdar Mahdi, Hassan, Masoud Muhammed, Mstafa, Ramadhan J., 2022. A review on deep sequential models for forecasting time series data. Appl. Comput. Intell. Soft Comput. 2022.

Ali, Ahmed, Fathalla, Ahmed, Salah, Ahmad, Bekhit, Mahmoud, Eldesouky, Esraa, 2021. Marine data prediction: an evaluation of machine learning, deep learning, and statistical predictive models. Comput. Intell. Neurosci. 2021.

Aparna, S.G., D'souza, Selrina, Arjun, N.B., 2018. Prediction of daily sea surface temperature using artificial neural networks. Int. J. Remote Sens. 39 (12), 4214–4231.

Bahdanau, Dzmitry, Cho, Kyunghyun, Bengio, Yoshua, 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Choi, Youngjin, Park, Youngmin, Hwang, Jaedong, Jeong, Kijune, Kim, Euihyun, 2022. Improving ocean forecasting using deep learning and numerical model integration. J. Mar. Sci. Eng. 10 (4), 450.

de Mattos, Paulo S.G., Neto, George D.C., Cavalcanti, Domingos S., de O Santos Júnior, and Eraylson G Silva., 2022. Hybrid systems using residual modeling for sea surface temperature forecasting. Sci. Rep. 12 (1), 1–16.

Fei, Tonghan, Huang, Binghu, Wang, Xiang, Zhu, Junxing, Chen, Yan, Wang, Huizan, Zhang, Weimin, 2022. A hybrid deep learning model for the bias correction of sst numerical forecast products using satellite data. Remote Sens. 14 (6), 1339.

Galassi, Andrea, Lippi, Marco, Torroni, Paolo, 2020. Attention in natural language processing. IEEE Trans. Neural Netw. Learn. Syst. 32 (10), 4291–4308.

Glorot, Xavier, Bengio, Yoshua, 2010. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 249–256. JMLR Workshop and Conference Proceedings.

Graves, Alex, Schmidhuber, Jürgen, 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Netw. 18 (5–6), 602–610.

Graves, Alex, Fernández, Santiago, Schmidhuber, Jürgen, 2005. Bidirectional lstm networks for improved phoneme classification and recognition. In: International Conference on Artificial Neural Networks. Springer, pp. 799–804.

Hochreiter, Sepp, Schmidhuber, Jürgen, 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

Hou, Siyun, Li, Wengen, Liu, Tianying, Zhou, Shuigeng, Guan, Jihong, Qin, Rufu, Wang, Zhenfeng, 2022. Mimo: a unified spatio-temporal model for multi-scale sea surface temperature prediction. Remote Sens. 14 (10), 2371.

Kingma, Diederik P., Ba, Jimmy, 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Krishnamurti, T.N., Chakraborty, Arindam, Krishnamurti, Ruby, Dewar, William K., Clayson, Carol Anne, 2006. Seasonal prediction of sea surface temperature anomalies using a suite of 13 coupled atmosphere–ocean models. J. Clim. 19 (23), 6069–6088.

Kun, Xiao, Shan, Tian, Yi, Tan, Chao, Chen, 2021. Attention-based long short-term memory network temperature prediction model. In: 2021 7th International Conference on Condition Monitoring of Machinery in Non-Stationary Operations (CMMNO). IEEE, pp. 278–281.

Lee, Dong Eun, Chapman, David, Henderson, Naomi, Chen, Chen, Cane, Mark A., 2016. Multilevel vector autoregressive prediction of sea surface temperature in the north tropical Atlantic Ocean and the caribbean sea. Clim. Dyn. 47 (1), 95–106.

Lee, Ming-Che, Chang, Jia-Wei, Yeh, Sheng-Cheng, Chia, Tsorng-Lin, Liao, Jie-Shan, Chen, Xu-Ming, 2022. Applying attention-based bilstm and technical indicators in the design and performance analysis of stock trading strategies. Neural Comput. & Applic. 1–13.

Lins, Isis Didier, Araujo, Moacyr, Márcio das Chagas Moura, Marcus André Silva, and Enrique López Droguett., 2013. Prediction of sea surface temperature in the tropical Atlantic by support vector machines. Comput. Stat. Data Anal. 61 (0), 187–198.

Luong, Minh-Thang, Pham, Hieu, Manning, Christopher D., 2015. Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025.

Ma, Jie, Jia, Chengfeng, Yang, Xin, Cheng, Xiaochun, Li, Wenkai, Zhang, Chunwei, 2020. A data-driven approach for collision risk early warning in vessel encounter situations using attention-bilstm. IEEE Access 8 (0), 188771–188783.

Nie, Qingqing, Wan, Dingsheng, Wang, Rui, 2021. Cnn-bilstm water level prediction method with attention mechanism. In: Journal of Physics: Conference Series, 2078. IOP Publishing, p. 012032.

Niu, Zhaoyang, Zhong, Guoqiang, Hui, Yu, 2021. A review on the attention mechanism of deep learning. Neurocomputing 452 (0), 48–62.

Noor, Fahima, Haq, Sanaulla, Rakib, Mohammed, Ahmed, Tarik, Jamal, Zeeshan, Siam, Zakaria Shams, Hasan, Rubyat Tasnuva, Adnan, Mohammed Sarfaraz Gani, Dewan, Ashraf, Rahman, Rashedur M., 2022. Water level forecasting using spatiotemporal attention-based long short-term memory network. Water 14 (4), 612.

Qiao, Baiyou, Zhongqiang, Wu, Ma, Ling, Zhou, Yicheng, Sun, Yunjiao, 2023. Effective ensemble learning approach for sst field prediction using attention-based predrnn. Front. Comp. Sci. 17 (1), 171601.

Rehana, Shaik, 2019. River water temperature modelling under climate change using support vector regression. In: Hydrology in a Changing World. Springer, pp. 171–183.

Shi, Jiahao, Jie, Yu, Yang, Jinkun, Lingyu, Xu, Huan, Xu., 2022. Time series surface temperature prediction based on cyclic evolutionary network model for complex sea area. Futur. Internet 14 (3), 96.

Sit, Muhammed, Demiray, Bekir Z., Xiang, Zhongrun, Ewing, Gregory J., Sermet, Yusuf, Demir, Ibrahim, 2020. A comprehensive review of deep learning applications in hydrology and water resources. Water Sci. Technol. 82 (12), 2635–2670.

Tuel, Alexandre, Eltahir, Elfatih A.B., 2018. Seasonal precipitation forecast over Morocco. Water Resour. Res. 54 (11), 9118–9130.

Wang, Feng, Tax, David M.J., 2016. Survey on the attention based rnn model and its applications in computer vision. arXiv preprint arXiv:1601.06823.

Wang, Yequan, Huang, Minlie, Zhu, Xiaoyan, Zhao, Li, 2016. Attention-based lstm for aspect-level sentiment classification. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 606–615.

Wang, Xiaoliang, Wang, Lei, Zhang, Zhiwei, Chen, Kuo, Jin, Yingying, Yan, Yijun, Liu, Jingjing, 2022. Sparse data-extended fusion method for sea surface temperature prediction on the East China Sea. Appl. Sci. 12 (12), 5905.

Wolff, Stefan, O'Donncha, Fearghal, Chen, Bei, 2020. Statistical and machine learning ensemble modelling to forecast sea surface temperature. J. Mar. Syst. 208 (0), 103347.

Xiao, Changjiang, Chen, Nengcheng, Chuli, Hu, Wang, Ke, Zewei, Xu, Cai, Yaping, Lei, Xu, Chen, Zeqiang, Gong, Jianya, 2019. A spatiotemporal deep learning model for sea surface temperature field prediction using time-series satellite data. Environ. Model Softw. 120 (0), 104502.

Xie, Jiang, Zhang, Jiyuan, Jie, Yu, Lingyu, Xu., 2019. An adaptive scale sea surface temperature predicting method based on deep learning with attention mechanism. IEEE Geosci. Remote Sens. Lett. 17 (5), 740–744.

Xuan, Yu, Shi, Suixiang, Lingyu, Xu, Liu, Yaya, Miao, Qingsheng, Sun, Miao, 2020. A novel method for sea surface temperature prediction based on deep learning. Math. Probl. Eng. 2020.

Xue, Yan, Leetmaa, Ants, 2000. Forecasts of tropical pacific sst and sea level using a markov model. Geophys. Res. Lett. 27 (17), 2701–2704.

Yang, Mo, Wang, Jing, 2022. Adaptability of financial time series prediction based on bilstm. Procedia Comput. Sci. 199 (0), 18–25.

Yang, Zichao, Yang, Diyi, Dyer, Chris, He, Xiaodong, Smola, Alex, Hovy, Eduard, 2016. Hierarchical attention networks for document classification. In: Proceedings of the

2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1480–1489.
Zhang, Xiaoyu, Li, Yongqing, Frery, Alejandro C., Ren, Peng, 2021. Sea surface temperature prediction with memory graph convolutional networks. IEEE Geosci. Remote Sens. Lett. 19 (0), 1–5.
Zheng, Gang, Li, Xiaofeng, Zhang, Rong-Hua, Liu, Bin, 2020. Purely satellite data–driven deep learning forecast of complicated tropical instability waves. Sci. Adv. 6 (29) eaba1482.