Prediction of Sea Surface Temperature by Combining Interdimensional and Self-Attention with Neural Networks

Xing Guo 1,*, Jianghai He 2, Biao Wang 3 and Jiaji Wu 1

1 School of Electronic Engineering, Xidian University, Xi’an 710071, China
2 School of Artificial Intelligence, Xidian University, Xi’an 710071, China
3 Science and Technology on Electromagnetic Scattering Laboratory, Shanghai 200438, China
* Correspondence: wuji@mail.xidian.edu.cn

Abstract: Sea surface temperature (SST) is one of the most important and widely used physical parameters for oceanography and meteorology. To obtain SST, in addition to direct measurement, remote sensing, and numerical models, a variety of data-driven models have been developed with a wealth of SST data being accumulated. As oceans are comprehensive and complex dynamic systems, the distribution and variation of SST are affected by various factors. To overcome this challenge and improve the prediction accuracy, a multi-variable long short-term memory (LSTM) model is proposed which takes wind speed and air pressure at sea level together with SST as inputs. Furthermore, two attention mechanisms are introduced to optimize the model. An interdimensional attention strategy, which is similar to the positional encoding matrix, is utilized to focus on important historical moments of multi-dimensional input; a self-attention strategy is adopted to smooth the data during the training process. Forty-three-year monthly mean SST and meteorological data from the fifth-generation ECMWF (European Centre for Medium-Range Weather Forecasts) reanalysis (ERA5) are collected to train and test the model for the sea areas around China. The performance of the model is evaluated in terms of different statistical parameters, namely the coefficient of determination, root mean squared error, mean absolute error and mean average percentage error, with a range of 0.9138–0.991, 0.3928–0.8789, 0.3213–0.6803, and 0.1067–0.2336, respectively. The prediction results indicate that it is superior to the LSTM-only model and models taking SST only as input, and confirm that our model is promising for oceanography and meteorology investigation.

Keywords: sea surface temperature; mutual information; LSTM; self-attention; interdimensional attention

1. Introduction

Sea surface temperature (SST) is one of the most important and widely used parameters in the analysis of global climate change. It is also used as boundary conditions or assimilation information in the analysis of atmospheric circulation anomalies, atmospheric models, and sea–air coupled models [1]. In addition, SST constitutes important basic data for aquaculture industry environmental assurance [2].

Although observations of SST have a history of more than 200 years, it was not until 1853 when the Brussels International Conference on Nautical Meteorology decided to start the collection of global SST data and standardize the organization and analysis of SST data. In recent decades, SST observation has transitioned through bucket observation measurements, Engine Room Intake (ERI) observations, ship-sensing observations, and satellite remote-sensing observations [3]. The uneven spatial and temporal distribution of observations need to be solved to obtain long-term, accurate global SST information. For this purpose, the reanalysis takes advantage of data assimilation techniques to integrate SST data from various sources and types of observations with numerical forecast products [4]. A number of reanalysis products that provide accurate forecasts across broad spatial and temporal scales have been released. In recent years, there has been a large volume of...
published studies comparing the products from different aspects. In 2013, Baololqimuge summarized several commonly used SST observation methods and introduced four sets of SST data including the Hadley Centre Sea Ice and Sea Surface Temperature data set (HadISST), the International Comprehensive Ocean-Atmosphere Data Set (ICOADS), Extended Reconstructed Sea Surface Temperature (ERSST) and Optimum Interpolation Sea Surface Temperature Analysis (OISSTA) in detail [3]. In the same year, Jiang conducted a comparative statistical analysis of six different SST products [5]. In 2020, Wang compared the applicability of three sets of reanalysis data around China [6].

The traditional methods for predicting ocean elements are divided into three main categories: numerical models, artificial experience and statistical prediction [4]. Numerical models are obtained under parametric conditions for physical processes in the ocean and are more suited to large sea areas and long-term SST prediction [7]. Furthermore, the computational requirements for ocean simulations are immense and require high-performance computing (HPC) facilities [8]. Artificial empirical and statistical prediction methods are more affected by parameter settings and the degree of human cognition.

Over the years, with the development, launch, and application of a series of satellites for oceanology and meteorology, satellite data have been increasing in magnitude. Additionally, the improvements in ocean models together with increased computational capabilities led to a number of reanalysis data [8]. The accumulation of a large amount of data on marine environmental elements has laid the foundation for data-driven methods [9]. These models abandon the subjectivity of traditional machine learning that requires experts to design and create feature extractors through experiments; they automatically and objectively extract useful information from data. Thus, they bring new opportunities for intelligent analysis and mining of marine data.

SST prediction can essentially be regarded as a time-series regression problem. The traditional models for time-series prediction, such as autoregressive (AR), moving average (MA), the autoregressive integrated moving average model (ARIMA) and the regression model by machine learning, including support vector regression (SVR) [10], and multi-layer perceptron (MLP) have been widely used in SST prediction [11]. An atmospheric reflection Grey model was proposed to predict long-term SST [12]. More recent attention has focused on the deep learning models, which originated from the artificial neural network (ANN). In 2006, Hinton proposed the concept of deep learning, which promoted the implementation of a number of deep learning projects [13]. A recurrent neural network (RNN) is a deep model developed for modeling sequential data [14]. The RNN introduces hidden states to extract features from sequential data and convert them to outputs. Hochreiter proposed the long short-term memory (LSTM) model, which introduces the forgetting gate and the memory gate [15]. In 2017, Zhang first adopted the LSTM model to predict SST. In the same year, based on the convolutional LSTM (ConvLSTM) model [16], Xu proposed a sequence-to-sequence (Seq2Seq)-based regional sea level temperature anomaly prediction model [17]. In 2018, Yang introduced spatial information with the LSTM model to build a model for SST prediction, and applied it effectively in the SST data set of coastal China [18]. In 2019, Zhu applied the LSTM-RNN to SST prediction and constructed a model for SST time-series variation in the western Pacific sea area [19]. LSTM was applied to predict SST and high-water temperature occurrence [2]. The temporal convolutional network (TCN) was applied to obtain large-scale and long-term SST prediction [20]. LSTM was applied to short and midterm daily SST prediction for the Black Sea [21].

To date, a number of researchers have attempted to combine different models together to predict SST. A numerical model is combined with neural networks to predict site-specific SST [22]. In 2019, Xiao combined LSTM and AdaBoost for medium-and long-term SST prediction [23]. Later, to fully capture the information of SST across both space and time, the author combined the convolutional network with LSTM as the building block to predict SST [24]. He combined the seasonal-trend decomposition using loess (STL) and LSTM to predict SST [25]. Deep learning neural networks were combined with numerical estimators for daily, weekly, and monthly SST prediction [7]. To enhance the
performance, a hybrid system which combines machine learning modes using residual forecasting was developed [26]. Jahanbakht designed an ensemble of two staked DNNs that used air temperature and SST to predict SST [27]. To forecast multi-step-ahead SST, a hybrid empirical model and gated recurrent unit was proposed [28]. Accuracy comparable to existing state of the art can be achieved by combining automated feature extraction and machine learning models [8]. Pedro evaluated the accuracy and efficiency of many deep learning models for time-series forecasting and show LSTM and CNN are the best alternatives [29].

LSTM has some advantages in sequence modeling owing to its long-time memory function; it is relatively simple to implement and solves the problem of gradient disappearance and gradient explosion that exists in the long sequence training process. However, it has disadvantages in parallel processing and always takes longer to train. A transformer, based on attention mechanisms, was proposed, which is parallelized and can significantly reduce the model’s training time [30]. Li enhanced the locality and overcame the memory bottleneck on the transformer for a the time-series prediction problem [31]. Furthermore, an SST prediction model based on deep learning with an attention mechanism was proposed [32]. A transformer neural network based on self-attention was developed, which showed superior performance than other existing models, especially at larger time intervals [33]. The degrees of effect on the prediction result of the information at previous time steps differ; therefore, the addition of an attention mechanism can assign different levels of attention to the model enabling it to automatically handle the importance of different information [34].

Inspired by transformer’s self-attention and positional encoding, the main contributions of this work can be summarized as follows:

1. The determining factors affecting SST distribution and variation, in other words, the input of the LSTM prediction model, is selected by the correlation analysis of mutual information.
2. To focus on important historical moments and important variables, a special matrix, that is similar to the position coding matrix, is obtained by multiplying the multi-dimensional data by a weight matrix W (where W is obtained by network training). The input data are smoothed using a self-attention mechanism during the training process.
3. The remainder of this paper is organized as follows. Section 2 first presents the correlation analysis of SST and meteorological data based on mutual information, and then describes the proposed model combining LSTM with attention mechanism. The study area and data sets used, implementation detail, and experimental results are introduced in Section 3. Validation of the model and comparison of its performance with other models are presented in Section 4. Finally, Section 5 concludes this paper and outlines future plans.

2. Methodology

The ocean is a comprehensive and complex dynamic system, and many factors affect the distribution and variation of SST. In the process of multivariate time-series model building, when the dimensionality of the input variables increases to a certain degree, the accuracy of parameter estimation decreases, which significantly decreases the prediction accuracy of the model and generates a dimensional disaster. In addition, the number of learning samples required for training increases exponentially with the dimensionality, whereas in practice the samples available for training are often very limited. By contrast, the model input with an excessive number of irrelevant, redundant, or useless variables, tends to obscure the role of the important variables eventually leading to poor prediction results [35].

Therefore, to identify valid inputs for SST prediction models based on deep learning, it is necessary to analyze the correlation between SST and meteorological and marine factors that may affect SST distribution and variation. On the one hand, by analyzing the correlation between input variables and output variables, the relevant variables that contribute most to the model prediction can be identified. On the other hand, by analyzing
whether there is some type of dependency between the input variables, redundant variables can be eliminated.

The present study involves the overall research plan shown in Figure 1. First, we used reanalysis data to construct a database of marine environmental elements, including SST, pressure, wind speed, solar irradiation, latitude, and longitude. Then, we perform quality analysis and corresponding preprocessing according to the analysis result. Because SST is affected by several factors simultaneously, to build a deep learning model, we determine the effective input of the model by analyzing the correlation among the influencing factors based on mutual information. Then, a hybrid model combining LSTM and attention mechanism is introduced. Subsequently, we evaluate the accuracy of the model for the surrounding sea areas of China.

![Figure 1. General flow chart of SST prediction based on a hybrid model combining LSTM and attention.](image)

2.1. Correlation Analysis

In general, correlation is used to describe the closeness of the relationship between variables. Correlations include asymmetric causal and driving relationships, as well as symmetric correlations. Among the traditional statistical methods, the Pearson correlation coefficient, Spearman correlation coefficient, and Kendall correlation coefficient are commonly utilized [35]. The Pearson correlation coefficient is used to measure the degree of linear correlation between two variables and requires the corresponding variables to be bivariate normally distributed. The Spearman correlation coefficient is used to analyze a linear correlation using the rank order of two variables; it does not require the distribution of the original variables and is a nonparametric statistical method [36]. The Kendall correlation coefficient is an indicator used to reflect the correlation of categorical variables and is applicable to the case where both variables are ordered categorically.

Commonly applied methods of correlation analysis of multivariate data include Copula analysis, random forest, XGBoost, and mutual information analysis [37]. The definition of mutual information is derived from the concept of entropy in information theory, which is often also called information entropy or Shannon entropy. Entropy expresses the degree of uncertainty in the values of random variables in a numerical form, thus describing the magnitude of information content of variables.

Based on the definition of probability density of data, mutual information is a widely used method to describe the correlation of variables. This is because there is no special requirement for the distribution of data types, and it can be used for both linear and nonlinear correlation analysis [35].

The information entropy of discrete random variables is defined as

$$H(x) = - \sum_{i=1}^{N} p(x_i) \log(p(x_i))$$

(1)

where $N$ is the number of samples and $p(x_i)$ is the frequency of $x_i$ in the data sets.
The mutual information of variable $X$ and variable $Y$ is defined as

$$I(X, Y) = \int \int p_{XY}(x, y) \log \frac{p_{XY}(x, y)}{p_X(x)p_Y(y)} dx dy$$

where $p_{XY}(x, y)$ is the joint probability density of $X$ and $Y$, $p_X(x)$ and $p_Y(y)$ are the marginal probability density of $X$ and $Y$, respectively.

According to the definition, when two variables $X$ and $Y$ are independent of each other or completely unrelated, their mutual information equals to 0, which implies that there is no jointly owned information between the two variables. When $X$ and $Y$ are highly dependent on each other, the mutual information will be large.

In practical problems, the joint probability density of the variables $(X, Y)$ is usually not known, and the variables $X$ and $Y$ are generally discrete. Therefore, the histogram method is commonly used. It discretizes the values of continuous variables by dividing the bins in the range of variables, putting different values of variables into different bins, then counting their frequencies, and subsequently performing calculation using the formula of discrete information entropy. However, determining the range size of each bin is difficult, and it usually requires repeated calculations to obtain the optimal solution.

Another commonly used method is called k-nearest neighbor estimation, which was first proposed in 1987 [38]. In 2004, the mutual information calculation method for computing two continuous random variables was proposed [39].

$$I^{(1)}(X, Y) = \psi(k) - \langle \psi(n_x + 1) + \psi(n_y + 1) \rangle + \psi(N)$$

$$I^{(2)}(X, Y) = \psi(k) - \frac{1}{k} - \langle \psi(n_x) + \psi(n_y) \rangle + \psi(N)$$

where $\langle \rangle$ is the mean value symbol and $\psi$ is the Digamma function calculated by the following iterative formula

$$\psi(1) = -0.5772516$$

$$\psi(x + 1) = \psi(x) + \frac{1}{x}$$

The results obtained by the two calculation methods are similar in most cases. However, in general, the first method has smaller statistical errors and larger systematic errors, and the second method is more suitable for the calculation of high-dimensional mutual information quantity.

The calculation time of k-nearest neighbor mutual information estimation mainly depends on the sample size, while it is less affected by the dimensionality of variable. Moreover, in general, the smaller the value of $k$ is, the larger is the statistical error and the smaller is the systematic error. Usually, $k$ is taken as 3.

2.2. Model Architecture

LSTM has been widely used in SST time-series prediction. However, the LSTM network requires a long training time because of the lack of parallelization ability. Further, the degree of effect at different time steps on the prediction result are different and varies dynamically with time. This cannot be handled by using LSTM exclusively. Inspired by the attention mechanism used in natural language processing, we added the attention structure into our model to enable it to automatically focus on important historical moments and important variables.

As Figure 2 shows, the model consists of five components. In addition to the necessary input and output module, a multivariate LSTM module is applied to capture the feature information in the time-series data. Integrating multi-dimensional information itself is difficult, because it is impossible to determine which dimension plays a more important role on the results. In addition, the importance of information tends to fluctuate with the time steps. Therefore, it is crucial to solve the problem reasonably linking multi-dimensional input data together to retain useful data and eliminate interfering data. The coefficient matrix $W$ (green part) is determined in a way similar to positional encoding in the attention
mechanism. In the blue part, whether the data are true is questioned. A self-attention approach is used to observe the difference between the true and predicted values of the adjacent data; based on this, the data in the current time step is fine-tuned. The weight of the data in the current time step is adjusted to make it closer to the true value in the next iteration.

Figure 2. Structural diagram of the attention-based LSTM model.

2.2.1. Interdimensional Attention Strategy

One of the major advantages of a transformer network is that, for a single isolated data point, it not only mines that data point’s information but also integrates multiple data together through positional encoding to mine the information between data. This approach is most suitable for text processing tasks, where a certain association between contexts exists and words are encoded in a uniform manner. However, single-dimensional time-series prediction tasks cannot be realized through this method. In case of SST prediction, owing to the scarcity of data, specific time-stamped information is often erased and only time-course data are retained as a set of information for any consecutive 12 months, rather than a fixed set of information for each month. The variation trend of the data differs for different starting months, even showing totally opposite trends, as illustrated in Figure 3. Moreover, determining the degree of correlation between the data from January 2010 and January 2011 is difficult. Therefore, positional encoding is not possible with only one-dimensional time-series data.

Figure 3. Variation trend of SST in a sliding window.

Note that SST at a specific location can be affected by various factors, including the wind speed, air pressure, and solar radiation, to varying degrees, and the general prediction algorithm, which often uses only the temperature data, has significant limitations. SST at a specific place has an implicit relationship with the meteorological factors with a high probability, which can be described by the following equation

$$T_{pred} = k \cdot T + v \cdot u_{10} + \ldots$$

where the coefficients $k$, $v$, and so on are unknown weight vectors, and their values cannot be directly determined on the basis of experience. The main problem is the possible contradiction and inconsistency of importance between parameters. Moreover, the parameters may vary with time steps. For example, the weight of temperature may be set to 0.8 and
that of wind speed to 0.2 in January; however, in February, the temperature weight may become 0.7 and the wind speed weight may change to 0.3. Therefore, their values can only be obtained on the basis of the training of the neural network. For each time step, there is a matrix of corresponding coefficient matrix \( k (i, v) \). Each coefficient vector is shown in the following equation.

\[
k = [k_1, k_2, \cdots k_n]
\]

where \( n \) is the length of time series. Together, these vector coefficients form the \( W \) matrix shown in Figure 2.

In this way, during the network training process, the \( W \) matrix gradually reveals the implicit connections between these different dimensional data. This combining of data of different dimensions produces an effect similar to the position encoding matrix in a transformer.

### 2.2.2. Self-Attention Smoothing Strategy

The adverse effects of inevitable systematic errors (e.g., temperature measurement errors, local temperature anomalies, weather anomalies at the time of temperature measurement, and human causes), can be reduced by requiring each data point to be self-conscious, as illustrated in Figure 4.

![Figure 4. Self-attention smoothing strategy.](image)

As shown in Figure 5, to determine whether it is smooth and fits the simulated curve, we need to calculate the relationship between the data at \( T_i \), \( T_{i-1} \) and \( T_{i+1} \). In specific implementations, the degree of fitting is judged by the difference between the predicted and actual values at the current time step is whether around the difference at the preceding and following time step or not. If it is smooth and fits the curve well enough, then the data are more reliable and the corresponding parameter \( k \) or \( v \) is increased accordingly (blue circle in Figure 4). If it is not smooth or does not fit the curve well enough (green circle in Figure 4), the data may be abnormal and the corresponding parameter \( k \) or \( v \) is reduced by a factor of 10%. A suitable value can be found after several training iterations.

![Figure 5. Specific execution process of self-attention smoothing.](image)
2.3. Evaluation Metrics

The performance and reliability of the model are evaluated in terms of the statistics of the coefficient of determination ($R^2$), $RMSE$, mean absolute error ($MAE$), and mean average percentage error ($MAPE$). They are defined as Equations (8)–(11), respectively. Here, $y_i$ represents the true SST values, $\hat{y}_i$ represents the predicted SST values, $m$ is the length of the test data sets, and $\bar{y}_i$ is the mean value of the true SST.

\[
R^2 = 1 - \frac{\sum_{i=1}^{m} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{m} (\bar{y}_i - y_i)^2} \tag{8}
\]

$R^2$ is in the range $[0, 1]$; 0 indicates that the model is poorly fitted, while 1 indicates that the model is error free. In general, the larger the $R^2$ is, the better the model.

\[
RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2} \tag{9}
\]

\[
MAE = \frac{1}{m} \sum_{i=1}^{m} |\hat{y}_i - y_i| \tag{10}
\]

\[
MAPE = \frac{100\%}{m} \sum_{i=1}^{m} \left| \frac{\hat{y}_i - y_i}{y_i} \right| \tag{11}
\]

The $RMSE$, $MAE$ and $MAPE$ range in $[0, +\infty)$; 0 indicates that the predicted value exactly matches the true value, and the larger the error, the larger the value.

3. Model Implementation and Experiment Results

3.1. Study Area and Data Sets

The study area focuses on the sea areas around China (Figure 6); the specific locations and representative characteristics are shown in Table 1. The distribution and variation of SST depend on multiple meteorological elements. For example, solar radiation has a heating effect on the sea surface. Wind is the direct driver of the upper ocean circulation, which is an important factor to determine the flow of the upper layer and affects the distribution of SST [19].

This study considers the temporal and spatial resolution of the data and the completeness of the environmental variables fully, then the fifth-generation ECMWF (European Centre for Medium-Range Weather Forecasts) reanalysis (ERA5) is selected to construct a multi-physical field data set of marine environmental elements for the sea areas around China. ERA5 provides hourly, daily and monthly estimates for a large number of atmospheric, ocean-wave and land-surface quantities [40].

<table>
<thead>
<tr>
<th>ID</th>
<th>Ocean Region</th>
<th>Longitude (E°)</th>
<th>Latitude (N°)</th>
<th>Average Depth (m)</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bohai Sea and North Yellow Sea</td>
<td>119–125</td>
<td>37–41</td>
<td>18</td>
<td>Nearly closed</td>
</tr>
<tr>
<td>2</td>
<td>South Yellow Sea</td>
<td>119–125</td>
<td>31–37</td>
<td>44</td>
<td>Semi-closed</td>
</tr>
<tr>
<td>3</td>
<td>East China Sea</td>
<td>121–125</td>
<td>29–31</td>
<td>370</td>
<td>Marginal sea</td>
</tr>
<tr>
<td>4</td>
<td>Taiwan Strait</td>
<td>119–121</td>
<td>24–25</td>
<td>60</td>
<td>Narrow strait</td>
</tr>
<tr>
<td>5</td>
<td>South China Sea</td>
<td>106–125</td>
<td>5–21</td>
<td>1212</td>
<td>Open sea area</td>
</tr>
</tbody>
</table>
Figure 6. Location of the seven sea areas around China.

The temporal resolution of the data used in this study is monthly and the data sets cover the period 1979–2021. The spatial resolution in latitude and longitude is 0.25°. As illustrated in Table 2, in addition to SST data, data on meteorological factors including wind speed, sea surface pressure, and sea surface solar radiation have been collected.

Table 2. Variables affecting SST distribution and variation.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Name</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>SST</td>
<td>Sea surface temperature</td>
<td>K</td>
</tr>
<tr>
<td>u10</td>
<td>Eastward component of the 10 m wind</td>
<td>m/s</td>
</tr>
<tr>
<td>v10</td>
<td>Northward component of the 10 m wind</td>
<td>m/s</td>
</tr>
<tr>
<td>msl</td>
<td>Mean sea level pressure</td>
<td>Pa</td>
</tr>
<tr>
<td>ssr</td>
<td>Surface net solar radiation</td>
<td>J/m²</td>
</tr>
<tr>
<td>ssrcc</td>
<td>Surface net solar radiation clear sky</td>
<td>J/m²</td>
</tr>
<tr>
<td>str</td>
<td>Surface net thermal radiation</td>
<td>J/m²</td>
</tr>
<tr>
<td>strcc</td>
<td>Surface net thermal radiation clear sky</td>
<td>J/m²</td>
</tr>
<tr>
<td>ssrd</td>
<td>Surface solar radiation downward</td>
<td>J/m²</td>
</tr>
<tr>
<td>ssrdcc</td>
<td>Surface solar radiation downward clear sky</td>
<td>J/m²</td>
</tr>
<tr>
<td>strd</td>
<td>Surface solar radiation downwards</td>
<td>J/m²</td>
</tr>
<tr>
<td>strdcc</td>
<td>Surface thermal radiation downward clear sky</td>
<td>J/m²</td>
</tr>
</tbody>
</table>

3.2. Implementation Detail

As shown in Table 2, the data sets consist of various parameters that have different units and ranges of values; thus, data normalization is necessary. The min-max normalization is utilized to scale the data between 0 and 1,

\[ z_i = \frac{x_i - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \]  

(12)

where \( x_i \) is the original data, \( x_{\text{min}} \) and \( x_{\text{max}} \) are the minimum and maximum of the original data, respectively.

Then, to transform the time series into input–output pairs required for model training, a sliding window with a fixed length is used as shown in Figure 7. In this study, the model receives an instance with a sliding window of length 12 as input and performs one-step predictions. The resulting samples are divided into training, validation, and test sets in a ratio of 6:1:2.
3.2. Implementation Detail

As shown in Table 2, the data sets consist of various parameters that have different units and ranges of values; thus, data normalization is necessary. The min-max normalization is utilized to scale the data between 0 and 1,

\[ \frac{x - \min_x}{\max_x - \min_x} \]

where \( x \) is the original data, \( \min_x \) and \( \max_x \) are the minimum and maximum of the original data, respectively.

Then, to transform the time series into input–output pairs required for model training, a sliding window with a fixed length is used as shown in Figure 7. In this study, the model receives an instance with a sliding window of length 12 as input and performs one-step predictions. The resulting samples are divided into training, validation, and test sets in a ratio of 6:1:2.

![Figure 7. Sliding window procedure to obtain input–output pair of data sets.](image)

Table 3. Key parameters of the model and training process.

<table>
<thead>
<tr>
<th>Key Parameters</th>
<th>Model Methods or Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length of training data sets</td>
<td>300</td>
</tr>
<tr>
<td>Length of validation data sets</td>
<td>50</td>
</tr>
<tr>
<td>Length of testing data sets</td>
<td>100</td>
</tr>
<tr>
<td>Architecture of the model</td>
<td>Attention + LSTM + Dense</td>
</tr>
<tr>
<td>Input dimension</td>
<td>12 × 4</td>
</tr>
<tr>
<td>Output dimension</td>
<td>1</td>
</tr>
<tr>
<td>No. of neural of hidden layer</td>
<td>80</td>
</tr>
<tr>
<td>Optimizer</td>
<td>Adam</td>
</tr>
<tr>
<td>Epoch</td>
<td>400</td>
</tr>
<tr>
<td>Batch size</td>
<td>40</td>
</tr>
<tr>
<td>Dropout</td>
<td>0.1</td>
</tr>
<tr>
<td>Loss function</td>
<td>RMSE</td>
</tr>
</tbody>
</table>

All experiments are implemented using Keras 2.2.4 with TensorFlow 1.15.0 on a computer with an Intel i9-10900K CPU and an additional NVIDIA GeForce RTX 2080S GPU.

3.3. Experiment Results

3.3.1. SST Distribution and Variation

From Figures 8–10, it can be concluded that the latitudinal distribution of SST is obvious, i.e., the South China Sea has a lower latitude and a higher temperature all year round. The annual variation, except for the South China Sea, shows a pattern of synchronous change with the temperature, i.e., the highest in August and the lowest in January.

3.3.2. Correlation of SST with Other Meteorological Factors

In this study, mutual information is selected as a tool to analyze the correlations between different environmental factors and SST, which is further required for selecting the effective input variable for building a deep learning prediction model.

The mutual information of SST with each influencing factor was calculated using k-nearest neighbor-based mutual information, as shown in Figure 11. The figure indicates that, overall, the wind speed (u10, v10) and air pressure at sea level (msl) correlate more strongly with SST in different seas compared to radiation-related environmental variables.
3.3. Experiment Results

3.3.1. SST Distribution and Variation

From Figures 8–10, it can be concluded that the latitudinal distribution of SST is obvious, i.e., the South China Sea has a lower latitude and a higher temperature all year round. The annual variation, except for the South China Sea, shows a pattern of synchronous change with the temperature, i.e., the highest in August and the lowest in January.

Figure 8. SST distribution of sea areas surrounding China in (a) March, (b) June, (c) September, and (d) December in 2021.

Figure 9. Variation of monthly mean SST for different sea areas around China during 1979–2021.

Figure 10. Statistics of monthly mean SST including minimum, maximum, and mean value for different parts of China’s Surrounding Seas during 1979–2021.

3.3.2. Correlation of SST with Other Meteorological Factors

In this study, mutual information is selected as a tool to analyze the correlations between different environmental factors and SST, which is further required for selecting the effective input variable for building a deep learning prediction model.

The mutual information of SST with each influencing factor was calculated using k-nearest neighbor-based mutual information, as shown in Figure 11. The figure indicates that, overall, the wind speed (u10, v10) and air pressure at sea level (msl) correlate more strongly with SST in different seas compared to radiation-related environmental variables.

Figure 9. Variation of monthly mean SST for different sea areas around China during 1979–2021.

Figure 10. Statistics of monthly mean SST including minimum, maximum, and mean value for different parts of China’s Surrounding Seas during 1979–2021.

Figure 11. Mutual information of SST with various meteorological factors.
3.3.2. Correlation of SST with Other Meteorological Factors

In this study, mutual information is selected as a tool to analyze the correlations between different environmental factors and SST, which is further required for selecting the effective input variable for building a deep learning prediction model.

The mutual information of SST with each influencing factor was calculated using k-nearest neighbor-based mutual information, as shown in Figure 11. The figure indicates that, overall, the wind speed (u10, v10) and air pressure at sea level (msl) correlate more strongly with SST in different seas compared to radiation-related environmental variables.

Figure 11. Heat map of mutual information for each meteorological variable and SST for the seas around China in different months (a) Bohai Sea and north part of Yellow Sea, (b) South part of Yellow Sea, (c) Part 1 (ID:3 in Table 1) of East China Sea, (d) Part 2 (ID:4 in Table 1) of East China Sea, (e) Part 1 (ID:5 in Table 1) of Taiwan Strait, (f) Part 2 (ID:5 in Table 1) of Taiwan Strait, and (g) South China Sea.

3.3.3. SST Prediction Results

The last 100 samples (about 8 years from 2014 to 2021) in the data sets are applied to test the model. The one-month ahead monthly mean SST prediction results of the sea areas around China are shown in Figure 12. The blue line represents the true values.
3.3.3. SST Prediction Results

The last 100 samples (about 8 years from 2014 to 2021) in the data sets are applied to test the model. The one-month ahead monthly mean SST prediction results of the sea areas around China are shown in Figure 12. The blue line represents the true values. Additionally, the red dot and the filled areas in the figure represent the average prediction results and the corresponding standard deviation for five runs. What should be noted is that the resolution of y axis is different for different regions.

Overall, the prediction results reveal a same trend between the true and predicted SST. However, for all regions, larger bias appears at the local extremums, because the model trained on the training data sets cannot capture the extremums of the test data sets. As SST of southern part of China, especially South China Sea, keeps high (approximately 300 K) all year round and fluctuates less, the model performs better.

To test the stability of the model, statistics including $R^2$, RMSE, MAE, and MAPE for five runs are presented in Table 4 and Figure 13. From the perspective of RMSE, MAE and MAPE, the model performs better in the southern parts of the surrounding seas of China, especially the South China Sea, for which the SST varies less and maintains a high value all year round. The error of some isolated point probably leads to higher RMSE, MAE and MAPE. The fluctuation for region 5 (Taiwan Strait) is the smallest, which may indicate that, for narrow strait areas, we can trust the result more from arbitrary initialization conditions.

Table 4. Comparison of $R^2$, RMSE, MAE and MAPE for seven study areas (average for five runs).

<table>
<thead>
<tr>
<th>Region ID</th>
<th>$R^2$</th>
<th>RMSE</th>
<th>MAE</th>
<th>MAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.9910</td>
<td>0.7551</td>
<td>0.6211</td>
<td>0.2170</td>
</tr>
<tr>
<td>2</td>
<td>0.9829</td>
<td>0.8789</td>
<td>0.6803</td>
<td>0.2336</td>
</tr>
<tr>
<td>3</td>
<td>0.9827</td>
<td>0.7547</td>
<td>0.5936</td>
<td>0.2029</td>
</tr>
<tr>
<td>4</td>
<td>0.9827</td>
<td>0.5120</td>
<td>0.4100</td>
<td>0.1382</td>
</tr>
<tr>
<td>5</td>
<td>0.9727</td>
<td>0.6515</td>
<td>0.5065</td>
<td>0.1711</td>
</tr>
<tr>
<td>6</td>
<td>0.9649</td>
<td>0.5666</td>
<td>0.4531</td>
<td>0.1521</td>
</tr>
<tr>
<td>7</td>
<td>0.9138</td>
<td>0.3928</td>
<td>0.3213</td>
<td>0.1067</td>
</tr>
</tbody>
</table>
and the corresponding standard deviation for five runs. What should be noted is that the resolution of y axis is different for different regions.

Overall, the prediction results reveal a same trend between the true and predicted SST. However, for all regions, larger bias appears at the local extremums, because the model trained on the training data sets cannot capture the extremums of the test data sets. As SST of southern part of China, especially South China Sea, keeps high (approximately 300 K) all year round and fluctuates less, the model performs better.

Figure 12. Cont.
Figure 12. Cont.
Figure 12. SST prediction results of the sea areas around China (a) Bohai Sea and north part of Yellow Sea, (b) South part of Yellow Sea, (c) Part 1 (ID:3 in Table 1) of East China Sea, (d) Part 2 (ID:4 in Table 1) of East China Sea, (e) Part 1(ID:5 in Table 1) of Taiwan Strait, (f) Part 2(ID:5 in Table 1) of Taiwan Strait and (g) South China Sea.

To test the stability of the model, statistics including $R^2$, RMSE, MAE, and MAPE for five runs are presented in Table 4 and Figure 13. From the perspective of RMSE, MAE and MAPE, the model performs better in the southern parts of the surrounding seas of China, especially the South China Sea, for which the SST varies less and maintains a high value all year round. The error of some isolated point probably leads to higher RMSE, MAE and MAPE. The fluctuation for region 5 (Taiwan Strait) is the smallest, which may indicate that, for narrow strait areas, we can trust the result more from arbitrary initialization conditions.

Figure 13. The distribution of the evaluation index for five runs: (a) RMSE, (b) R-squared, (c) MAE and (d) MAPE.
4. Discussion

4.1. Performance Comparison with Other Models

Table 5 shows the performance comparison of two other models with the model considering attention mechanism and taking both SST and meteorological factors as inputs. One of the models is the LSTM model taking SST only as input, and the other model takes SST and meteorological factors as inputs without considering attention mechanism. The boldface items in the table represent the best performance. The hyper-parameters affecting the training process are the same for the models.

Table 5. Comparison of RMSE between different models.

<table>
<thead>
<tr>
<th>Region ID</th>
<th>LSTM Only with SST Only as Input</th>
<th>LSTM Only</th>
<th>Our Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0157</td>
<td>0.9226</td>
<td>0.7551</td>
</tr>
<tr>
<td>2</td>
<td>1.0302</td>
<td>0.8657</td>
<td>0.8789</td>
</tr>
<tr>
<td>3</td>
<td>1.0481</td>
<td>0.7853</td>
<td>0.8789</td>
</tr>
<tr>
<td>4</td>
<td>0.8388</td>
<td>0.7301</td>
<td>0.7547</td>
</tr>
<tr>
<td>5</td>
<td>1.2139</td>
<td>0.8768</td>
<td>0.5120</td>
</tr>
<tr>
<td>6</td>
<td>0.7678</td>
<td>0.6487</td>
<td>0.5666</td>
</tr>
<tr>
<td>7</td>
<td>0.4140</td>
<td>0.4018</td>
<td>0.3928</td>
</tr>
</tbody>
</table>

It can be seen from the results that our model achieves the best performance for most regions. For the South China sea areas, three models show similar performance. Thus, it enables researchers to use the simple LSTM-only model with SST only as input for predicting SST in southern regions of China when there are insufficient meteorological data or computing resources.

4.2. Overfitting Issue Verification

To test if the trained model has overfitting issue, we have done another experiment to validate the generalization capability of the model. The forty-three-year (1979–2021) monthly mean SST and meteorological time-series data from ERA5 are used to train and validate the model. Then, the eight-year (1971–1978) data sets are fed into the trained model.

The prediction results shown in Figure 14 are the average for five runs, which verify the applicability and effectiveness of model. The black and red line represent the true values average prediction results.

![Figure 14](image-url)
Figure 14. Cont.
5. Conclusions

SST is a significant physical parameter used in the analysis of the ocean and climate. This study developed a data-driven model for predicting one-month ahead monthly mean SST by combining interdimensional and self-attention mechanism with neural networks. After correlation analysis by mutual information, SST and other meteorological factors including wind speed and air pressure were selected as the input of the prediction model. The interdimensional attention enabled the model to focus on important historical moments and important variables while the self-attention mechanism was utilized to smooth the data in the training process. Forty-three-year monthly mean SST and meteorological time-series data from ERA5 of ECMWF were collected to train the model and test its performance for the sea areas around China. The evaluation criteria of $R^2$, RMSE, MAE and MAPE indicate that the predicted results met the requirement for oceanography and meteorology studies.

During experiment, we find that, in most cases, other meteorological factors contribute to the predicted results, but these data, especially the wind speed, are not as stable as SST data and are prone to anomalies. The model is unable to reduce its coefficients quickly enough, thus leading to a longer training process eventually.

Overall, the performance of the model on SST prediction is promising. Future work involves further optimization of the model and investigation of its applicability for other ocean physical parameters such as sea surface salinity, and ocean water temperature underneath the surface.
Author Contributions: Conceptualization, X.G. and J.H.; methodology, X.G. and J.H.; validation, X.G. and B.W.; investigation, X.G.; resources, B.W.; writing—original draft preparation, X.G.; writing—review and editing, X.G.; visualization, X.G. and J.H.; conceptualization and supervision, J.W.; funding acquisition, X.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (grants 62005205 and 62010297) and the Shaanxi Province Science Foundation for Youth, grant number 2020JQ-329.

Data Availability Statement: SST and meteorological data are from fifth-generation ECMWF reanalysis data ERA5. The data are open and freely available at https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels-monthly-means?tab=overview, accessed on 16 February 2022.

Acknowledgments: The authors are grateful to ECMWF for supporting data.

Conflicts of Interest: The authors declare no conflict of interest.

References


