

Sea Surface Temperature Prediction Approach Based on 3D CNN and LSTM with Attention Mechanism

Baiyou Qiao*, Zhongqiang Wu*, Zhong Tang*, Gang Wu**

*School of Computer Science and Engineering, Northeastern University, China

**Key Laboratory of Intelligent Computing in Medical Image, Ministry of Education, Northeastern University, China

qiaobaiyou@mail.neu.edu.cn, 17862328070@163.com, 2071805@stu.neu.edu.cn, wugang@mail.neu.edu.cn

Abstract— Sea Surface Temperature (SST) is an important physical quantity of the ocean system. The accurate prediction of SST is essential for studying physical ocean phenomena and forecasting the ocean environment information. In this paper, a SST prediction approach based on 3-Dimensional Convolutional Neural Network (3D CNN) and Long Short-Term Memory (LSTM) network with attention mechanism is proposed, which considers the spatial correlation and temporal dependency of SST data. Firstly, the machine learning algorithm XGBoost is used to extract the long period time features of each SST data. Then the 3D CNN is used to capture the spatial correlations among SST field data composed of multiple observation points in a selected sea area, followed by the LSTM model to extract the time dependency features of the SST field time series data, and the attention mechanism is added to weight the output of each step of LSTM model to adjust the prediction results and improve the prediction accuracy of the approach. A series of experimental results show that the proposed approach has lower complexity, higher training efficiency and prediction accuracy, which is significantly better than the existing prediction models.

Keywords— Sea surface temperature; XGBoost; 3D CNN; LSTM; Attention mechanism;

I. INTRODUCTION

Sea surface temperature (SST) is one of the most important parameters of the global ocean, which plays a fundamental role in exchange of energy, momentum and moisture between the oceans and atmosphere [1]. Sea surface temperature prediction is the key link of ocean environmental information prediction. It is widely used in ocean fishery [2], global warming [3], marine disaster prevention and other ocean related fields.

At present, many prediction methods have been proposed for SST prediction, which can be divided into two categories, numerical methods and data-driven methods [1]. Numeric methods make use of a series of complex physics equations to describe SST variation rules, which is usually very complicated and demand massive computational efforts [4,5,6]. Data-driven methods attempt to solve the SST prediction problem from a data-centric perspective, automatically learn SST variation rules from SST data and use the learned patterns to infer future SST. Data-driven methods include the time series analysis, regression analysis, Markov model [7], Support vector machine (SVM) [8], neural network [9], SparkDTW plus SVM [10] and etc. Most of these methods are mainly based on traditional

statistical and machine learning, which are suitable for processing small SST data, usually used for predicting SST of single observation point, and not suitable for predicting SST field composed of multiple observation points in a sea area.

In recent years, with the rapid development of deep learning techniques [11], researchers have begun to apply deep learning approaches to ocean data analysis and prediction. [12] proposed a SST prediction model based on LSTM, which can capture the temporal dependency relationship among SST data and improve the SST prediction precision. [6] established CFCC-LSTM model to predict SST, which was fused into a CNN to calculate the spatial correlation between points in the region on the basis of a RNN capturing temporal features. CFCC-LSTM has smaller prediction error. ConvLSTM model integrates convolutional operations with LSTM to predicts SST [1]. The model can capture spatial and temporal correlations of SST data and can obtain better prediction results. But the above approaches not take into account the importance of historical information, which leads to an increase in network complexity and a significant decrease in prediction accuracy and efficiency when the prediction step and historical information increase. In addition, the above-mentioned studies generally use the original data to train the models directly. [13] presents a separate training method (STM) focusing on the impact of training data input mode. Compared with the previous direct training method, STM aims at reducing the prediction error caused by SST fluctuations and thereby, improving the prediction. However, it only used for predicting SST of single observation point rather than a SST field composed of multiple observation points.

In this paper, we proposed a SST prediction approach based on 3D CNN and LSTM with attention mechanism (3DCNN-LSTM-AT), which not only consider the impact of training data input mode, but also take into account the importance of historical information. the approach applies XGBoost model to extract long period time features of SST data, uses 3D CNN model to capture spatial features among a SST field data composed of multiple observation points in a local sea area, and then combining with excellent temporal feature learning ability of LSTM, the model can effectively extract spatial and temporal dependence relationship among SST field data. However, with the increase of the prediction step, the error between each step is also increasing. To solve the problem, an

attention mechanism is introduced to correct the output of LSTM by learning appropriate weights on each step, which achieves the high-precision and long-term SST prediction. A series of experimental results show that the proposed approach can achieve better prediction accuracy in the short-term, medium-term and long-term SST predictions, especially in the medium-term and long-term SST predictions, our approach is significantly better than the existing SST prediction approaches.

The rest of this paper is organized as follows. Section 2 describes the prediction problem of the SST field. Section 3 presents the proposed SST field prediction approach. The experimental result and thorough discussion are given in Section 4. Finally, Section 5 concludes the paper.

II. DESCRIPTION OF SST FIELD PREDICTION PROBLEM

The goal of SST field prediction is to use the historic SST field data to forecast a fixed length of SST field. From the machine learning perspective, this problem can be regarded as a spatiotemporal sequence forecasting problem [14].

Suppose we observe a dynamical SST field over a spatial region represented by an $M \times N$ grid which consists of M rows and N columns. In each cell, there are V measurements which vary over time (as we only use SST as the predicting variable, the value of V is 1). Thus, the observation at any time can be represented by a tensor $X_t \in R^{M \times N \times V}$, where R denotes the domain of the observed features [1]. In a period of time n , we can obtain a spatiotemporal sequence of observations $X_{t-n+1}, X_{t-n+2}, \dots, X_t$. The spatiotemporal sequence prediction problem is thus to forecast the most likely tensor of length k in the future given the previous n observations. In SST forecasting, it can be divided into short-term, medium-term and long-term forecasts depending on the size of k . Thus, SST field prediction problem can be formulated in equation (1).

$$\tilde{X}_{t+1}, \dots, \tilde{X}_{t+k} = \underset{X_{t+1}, \dots, X_{t+k}}{\operatorname{argmax}} P(X_{t+1}, \dots, X_{t+k} | X_{t-n+1}, X_{t-n+2}, \dots, X_t) \quad (1)$$

where $P(\cdot)$ is a conditional probability and $\tilde{X}_{t+1}, \dots, \tilde{X}_{t+k}$ are the predicted SST fields at time $t+1, \dots, t+k$.

III. SST PREDICTION APPROACH BASED ON 3D CNN AND LSTM WITH ATTENTION MECHANISM

In order to solve the SST field prediction problem, a new prediction approach 3DCNN-LSTM-AT is proposed in this paper, which is based on 3D CNN and LSTM model with attention mechanism. It not only takes into account the spatial correlation of SST field data in the same time, but also consider the time dependence among SST field data at different time, it also uses attention mechanism to weight the output of each step of LSTM to improve the prediction results. Thus, it can predict the medium-term and long-term SST changes more accurately. The framework of 3DCNN-LSTM-AT approach is shown as Figure 1, which is mainly divided into four stages, that are long period time features extraction, spatial features extraction, temporal dependency learning, weighting and output results.

Long period time features extraction. Considering the seasonal periodicity of SST data, the Pearson correlation coefficient and XGboost model [15] are used to extract the long period time features of SST data, which together with the raw

SST data are used as inputs the 3D CNN model in next phase.

Spatial features extraction. In spatial features extraction phase, 3D CNN model is used to extract the spatial dependence features between the SST field data. 3D CNN is developed on the basis of 2D CNN, which can well extract the spatial features of the data and capture part of the time features.

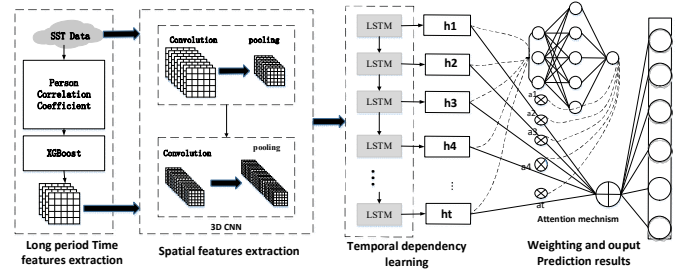


Figure 1. the SST prediction framework of 3DCNN-LSTM-AT

Temporal dependency learning. In this stage, the LSTM model is used to capture the time dependence among SST field data at different time.

Weighting and output prediction results. In this phase, the attention mechanism is used to weight and fine-tune the output of each step of the LSTM model, and then a fully connected network is used to obtain the final prediction results.

It can be seen that the 3DCNN-LSTM-AT approach takes into account both spatial and temporal correlations of SST field data to increase the prediction accuracy of the model. The following will be discussed in detail.

A. Long Period Time Features Extraction

From section 2, the goal of SST prediction is to use the previously observed SST sequence to forecast a fixed length of the SST in a region. Most of the existing algorithms generally turn the problem into supervised learning by using a sliding window method, where the historical SST data is partitioned into segments, and the former part of each segment data is used as the feature and the latter part as the label to train the model and make predictions. In this paper, considering the seasonal influence on the SST, the Pearson Correlation Coefficient method is used to select the four most relevant time features from the seven time features of the SST data, and then the XGBoost model is used to fuse them into a long period time feature, which will be used as the input for the next stage together with the SST data, so as to improve the accuracy of the prediction by increasing the new features of SST data.

Assuming that there are 64 observations at the sea surface in a local area, the historical temperature series for each observation $X_i, i \in \{1, 2, \dots, 64\}$ can be seen a time series. X_i is a vector that containing temperature information $X_{ij}, j \in \{1, 2, \dots, n\}$ for each period j of the observation point i, n is the length of time series. For each data X_{ij} , four long periodic time features are extracted, which is a tuple of four elements, the elements are quarter, month, week and season, denoted as $T_{ij} = \{\text{quarter}_{ij}, \text{week}_{ij}, \text{month}_{ij}, \text{season}_{ij}\}$. Using T_{ij} and the raw SST data as the inputs to the XGBoost model, the fused time feature can be obtained from XGBoost model for each SST data. The fused time feature can be seen as the long period time

feature of the SST, denoted by XT_{ij} , which represents the dependency relationship between the SST X_{ij} and the time feature T_{ij} .

Combining XT_{ij} with the X_{ij} to generate a new data $D_{ij} = \{X_{ij}, XT_{ij}\}$, which represents the new data features of the i -th observation point in period j . Based on the new data D_{ij} , under the sliding window, we can get n -periods historical data $X_{ij-n+1}, X_{ij-n+2}, \dots, X_{ij}$, it can be denoted by D_i . Merge all D_i and generate a data set D , which can be used as the input of the 3D CNN in next stage.

B. Space Features Extraction

In this phase, 3D CNN model is used to extract the spatial features among the SST field data at the same time through two convolution and pooling operations. The input of the 3D CNN is a 5-dimensional tensor expressed as $B \times F \times L \times W \times C$, where B is the number of training samples in a batch, F is the length of SST field series data. L and W are the length and width of the SST field. C is the channel information of the model, the value of C here is 1. In the following experiments, both length and width of SST field are set to 8, that means values of L and W are set to 8, and the numbers of features on past time is obtained by sliding the window. For example, if the SST is predicted by past 4 days SST data, then the sequence length is 4, in addition, 4 long-period time features extracted by XGBoost, the length of the time series is 8, i.e. the value of F . Finally, the data from each window are combined to obtain a 5-dimensional tensor as an input to the next model 3D CNN.

The input dimension of SST data X is $B \times 8 \times 8 \times 8 \times 1$. Next, the 3D CNN captures the spatial correlations of SST data, which is calculated by equation (2).

$$\begin{cases} X = \text{Conv3D}(X) \\ X = \text{MaxPooling3D}(X) \\ X = \text{Conv3D}(X) \\ X = \text{MaxPooling3D}(X) \end{cases} \quad (2)$$

Here, a two-layer convolution add pooling operation is used. Layer1 convolution operation uses 64 filters, the size of the filters is $3 \times 3 \times 3$. Pooling size is set to $2 \times 2 \times 2$. Layer2 convolution sets up 32 filters, kernel size is $3 \times 3 \times 3$. Pooling layer size is set to $2 \times 2 \times 2$. After two-layer convolution and pooling operation, the dimension of final data becomes $[-1, 8, 32]$. The output of 3D CNN model is the input of the next model LSTM. The 3D CNN is schematically layered as Figure 2.

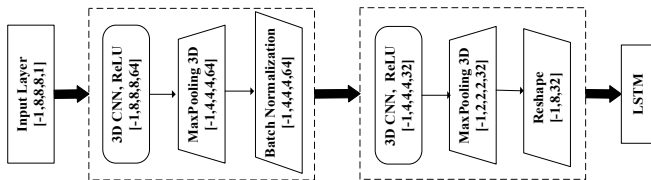


Figure 2. 3D CNN Layered Diagram

C. Temporal Dependence Learning

LSTM network model solves the problem of gradient vanishing of traditional RNN, and can capture long-distance dependence, which is very suitable for time series data feature extraction. Therefore, LSTM model is used to learn the time-

dependency features between the SST field data. The calculations of LSTM model are shown in equation (4).

$$\begin{aligned} i_t &= \text{sigmoid}(W_i h_{t-1} + U_i x_t + b_i) \\ f_t &= \text{sigmoid}(W_f h_{t-1} + U_f x_t + b_f) \\ o_t &= \text{sigmoid}(W_o h_{t-1} + U_o x_t + b_o) \\ \tilde{c}_t &= \tanh(W_c h_{t-1} + U_c x_t + b_c) \\ c_t &= f_t * c_{t-1} + i_t * \tilde{c}_t \\ h_t &= o_t * \tanh(c_t) \end{aligned} \quad (4)$$

where $W_i, W_o, W_f, W_c, U_i, U_o, U_f, U_c, b_i, b_f, b_o$, and b_c are the parameters of the LSTM model, these parameters are set to initial values, its values are adjusted through model training. The vector h_{t-1} represents the hidden layer output of the LSTM at $t-1$ th moment, and x_t denotes the input vector at t -th moment. 1 layer of LSTM is used here, the hidden layer unit is set to 15, the activation function is RELU, the formula is shown in equation (5).

$$H = \text{LSTM}(X) \quad (5)$$

After the LSTM, the dimension of the output result is $B \times 8 \times 15$, this is the input of the Attention layer. With the combination of 3D CNN and LSTM, the spatial and temporal features of the SST field time series data can be captured. The LSTM is followed by the attention layer, which learns a weight for each output that indicates the degree of correlation between the historical sequence and the current prediction.

D. Weighting and Output Prediction Results

Attention mechanism [16] [17] can assign different weights to different parts of the model, allowing the model to focus more on the parts that are more relevant to the task, thus improving the quality of model. In long-term SST prediction task, as the prediction step increases, more and more historical data are required for each prediction. If the data are directly input to the LSTM, it will increase the network's complexity, and the data are of different importance as to current prediction, so it should not be treated equally [18]. For this reason, a custom attention layer is added behind the LSTM network that can add appropriate weights on outputs of LSTM steps to adjust the finally LSTM output and thus obtain better prediction results. A custom attention layers is shown as Figure 3.

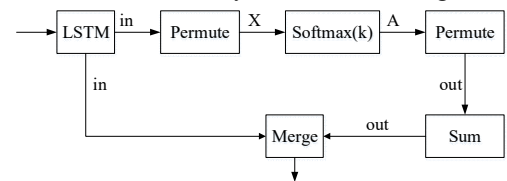


Figure 3. A custom attention layer

Permute is a transformation operation of the data dimension, the attention layer takes the output X of each iteration of LSTM as input, gets the weight of each output vector through *Softmax* operation, multiplies and sums the output of each iteration of LSTM network with its corresponding weight to get the weighted vector, and then maps it to the final output result through the full connection layer. The main formula is calculated as equation (6):

$$\begin{aligned}
K &= \tanh(X \cdot W + b) \\
A &= \text{softmax}(K) \\
out &= \text{sum}(A \times X)
\end{aligned} \tag{6}$$

Where X represents the output of LSTM network, the dimension of X is $B \times F \times H$, H represents the number of units in the hidden layer of LSTM. W represents the weight matrix of each output unit of LSTM in the feature dimension, A represents the weight of each output unit of LSTM in the feature dimension, out represents the final weight vector of each output unit.

The input data dimension of the Attention layer is $B \times 8 \times 15$ and the output dimension is still $B \times 8 \times 15$, the difference between the input and output is that the latter has a weight of the time dimension. A fully connected layer is then used to obtain the final prediction results. the layering of the LSTM and Attention layers is shown Figure 4.

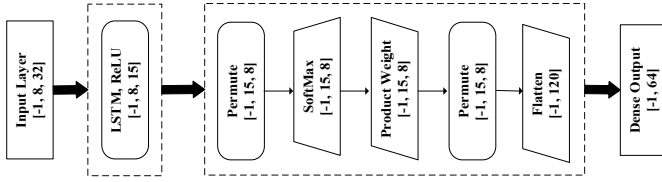


Figure 4. LSTM Attention Hierarchy Diagram

IV. EXPERMENTS AND RESULTS

A. Experimental Environment and Datasets

The experimental environment consists of a server with E5-2680 V4 CPU (14 cores, 2.4 GHz, 64GB memory, 4TB disk, one NVIDIA RTX 2080 TI GPU). The machine is set up to run the CentOS-6.4 operating system, Keras, TensorFlow 1.5.0 and the related algorithm modules implemented by Python.

The datasets used in this experiment are high resolution SST data from website <http://www.esrl.noaa.gov/psd/>. The datasets contain SST daily mean, SST weekly mean, and SST monthly mean from September 1, 1981 to present and cover the global ocean range 98.875S-98.875N, 0.125E-359.875E, and they are in the form of a $0.25^\circ\text{C} \times 0.25^\circ\text{C}$ grid. From the above data sets, the daily mean SST data of an 8×8 region in Bohai and South China Sea are selected for model training and testing respectively. The latitude and longitude range of Bohai Sea SST dataset is 37.875N-39.625N, 110.125E-120.875E; the longitude and latitude range of South China Sea dataset is 4N-6.25E, 105E-107.25E. Time range of the data sets is from January 1, 1982 to December 31, 2018, with a total of 13514 SST field time series data.

B. Parameter Settings

We tested the proposed 3DCNN-LSTM-AT approach and compared it with LSTM and ConvLSTM prediction approaches. The parameters of these models are as follows.

For 3DCNN-LSTM-AT approach, the number of iterations of the XGboost is 500, learning rate is 0.07, the maximum number of training rounds of 3DCNN-LSTM is set to 200, the optimizer is Adam, the learning rate is decayed, the initial learning rate is 0.001, the decay coefficient is 0.2 or 0.1 and the ratio of model training set to verification set is 0.03.

The first layer of ConvLSTM network has 128 filters, size is 5×5 , layer2 has 64 filters, size is 5×5 . And then a full connection layer where the number of neural units is the product of the output dimensions and the prediction step size. The LSTM prediction model is a one-layer network, its number of hidden layer units is 15. And the next is full connection layer. The number of neural units is the product of output dimension and prediction step size. The ratio of training set to verification set is 0.03.

C. Experimental Results

For the short-term, medium-term, and long-term prediction tasks, we used 13514 SST field time series data from the Bohai Sea and the South China Sea, respectively. The first 13,149 SST data are used as training set and the last 365 SST data are used as test set. In order to facilitate analysis and comparison, the 3DCNN-LSTM model proposed in this paper is divided into four situations, with no long period time feature and attention mechanism (3DCNN-LSTM), only long period time features (3DCNN-LSTM-TF), only attention mechanism (3DCNN-LSTM-AT-ONLY), long period time feature and attention mechanism (3DCNN-LSTM-AT), and tested respectively. Four indicators, Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Coefficient of Determination (R^2) are used to evaluate these models. The following will give the detail experimental results and analysis from short-term prediction, medium-term prediction and long-term prediction three aspects.

1) Short-term Prediction: The short-term prediction mainly uses the SST data in past 4 days to predict the SST of the next day. Figure 5 shows the comparison of prediction errors of several models on South China Sea dataset. It can be seen from the figure that 3DCNN-LSTM-AT has the minimum prediction error and the highest R^2 score, and the fitting ability of the model is better than other models. When no long period time feature is added, the prediction effect of 3DCNN-LSTM and ConvLSTM is not much different, because both of them can capture the spatial and temporal dependence features of the SST sequences. The difference is that 3DCNN-LSTM learns the space features first and then the temporal features, while ConvLSTM learns both temporal and spatial features at the same time. Adding long period time features and attention mechanism is helpful to improve the prediction accuracy. LSTM is relatively poor, because LSTM model only extracts the temporal features of SST data, but cannot extract the spatial features of SST data.

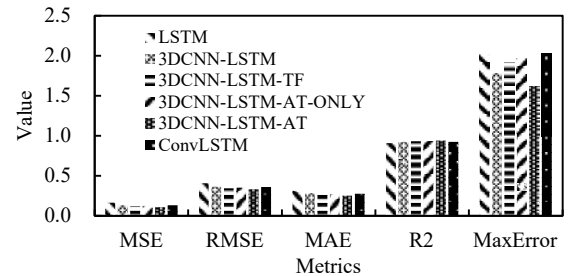


Figure 5. Comparison of short-term forecasting model errors

2) Medium-term Prediction: The medium-term prediction uses the past 12 days SSTs to predict the SSTs of the next 3 days. With the increase of the prediction step size, the longer the past time series data needed will be. Figure 6 shows the variation of MSEs of several models with the prediction step size on Beihai dataset. It is obviously seen from the figure that with the increase of prediction step size, the MSE of each model shows an upward trend, which is consistent with the expectation. The 3DCNN-LSTM-AT approach has the lowest MSE, achieving better fitting effect. Without time features and attention mechanism, the prediction accuracy of 3DCNN-LSTM and ConvLSTM is still not much different. Table 1 shows the scores of RMSE, MAE, and R^2 of the several models on two data sets in Bohai Sea and South China Sea region.

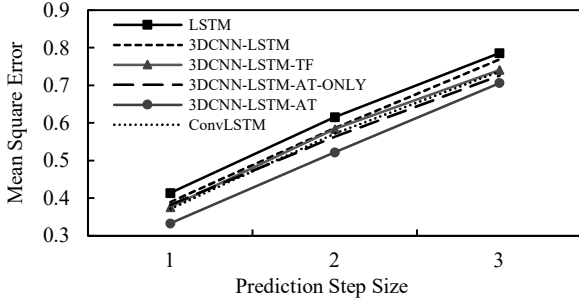


Figure 6. Comparison of medium-term forecasting model errors

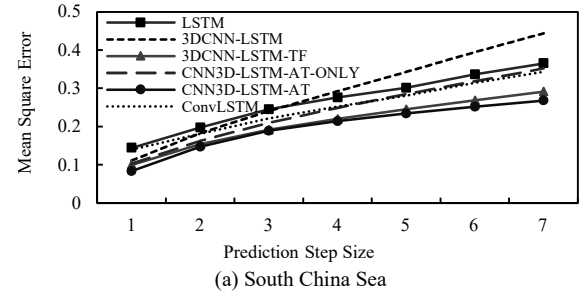
Table 1. The comparisons of models in Bohai Sea (South Sea) data sets

Model	Step	RMSE	MAE	R^2
LSTM	1	0.643(0.391)	0.466(0.299)	0.978(0.911)
	2	0.784(0.437)	0.575(0.337)	0.976(0.889)
	3	0.886(0.472)	0.645(0.366)	0.973(0.869)
3DCNN-LSTM	1	0.624(0.379)	0.458(0.291)	0.979(0.916)
	2	0.766(0.436)	0.565(0.338)	0.976(0.889)
	3	0.877(0.490)	0.640(0.382)	0.973(0.859)
3DCNN-LSTM-TF	1	0.613(0.345)	0.448(0.263)	0.979(0.930)
	2	0.764(0.410)	0.568(0.315)	0.976(0.901)
	3	0.861(0.448)	0.638(0.347)	0.974(0.882)
3DCNN-LSTM-AT-ONLY	1	0.618(0.355)	0.449(0.274)	0.973(0.926)
	2	0.750(0.416)	0.551(0.323)	0.976(0.898)
	3	0.852(0.466)	0.620(0.365)	0.974(0.872)
3DCNN-LSTM-AT	1	0.577(0.349)	0.412(0.268)	0.979(0.929)
	2	0.723(0.404)	0.538(0.313)	0.977(0.904)
	3	0.8406(0.4429)	0.6192(0.3446)	0.974(0.885)
ConvLSTM	1	0.609(0.376)	0.441(0.289)	0.979(0.917)
	2	0.756(0.450)	0.551(0.349)	0.976(0.881)
	3	0.858(0.504)	0.623(0.393)	0.974(0.851)

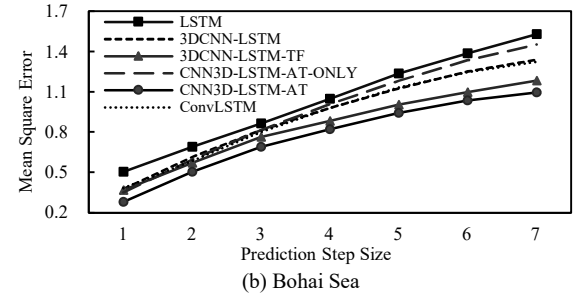
Taken together, 3DCNN-LSTM-AT model has better prediction accuracy in medium-term SST prediction.

3) long-term Prediction: The long-term prediction uses the SST data of the past 21 days to predict the SST in the next 7 days, for the SST field is composed of 64 points, so the input and output dimensions are relatively high. For the network models, the training speed is very slow, limited by the memory of GPU accelerator. In the long-term prediction, the cycle prediction strategy is adopted, that is, the SST data of 21 days is used to predict the SST of the next day, and then the predicted SST is put into the test set to predict the SST of the next day, thus cycle to predict the SST in the next 7 days. Figure 7 shows the long-term prediction results of several models. Figure 7 (a) and (b) denote the MSE of each model on two datasets,

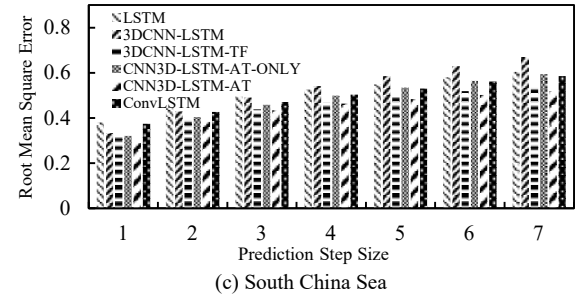
respectively. Figure 7(c) and (d) represent the RMSE of two datasets. As can be seen from the figures, with the increase of prediction step size, the prediction error of several models gradually increases. The 3DCNN-LSTM-AT model maintains the minimum prediction error on both datasets, and the fluctuation between step sizes is small. Therefore, time features and attention mechanism are conducive to improving the accuracy of the model. The performance of LSTM in Bohai Sea data set is the worst. Only the 3DCNN-LSTM model of attention mechanism has a small initial error. With the increase of step size, the error becomes larger and larger, so the fluctuation is relatively large. The same fluctuation of ConvLSTM is also relatively large. The reason may be that the fluctuation of Bohai Sea data itself is very large, which adds some challenges to the prediction models. The data set of the South China Sea is relatively stable.



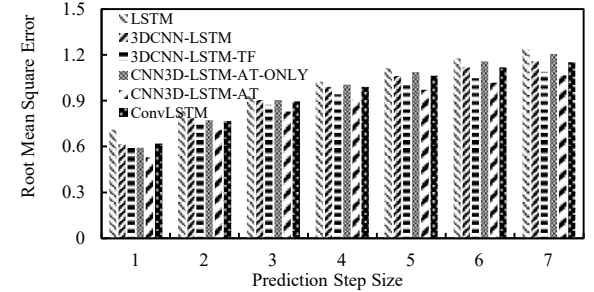
(a) South China Sea



(b) Bohai Sea



(c) South China Sea



(d) Bohai Sea

Figure 7. Comparison of long-term forecasting of the several models

From comparative experiments of short-term prediction, medium-term prediction and long-term prediction, we can see that the proposed 3DCNN-LSTM-AT model, which combines temporal features and attention mechanism, has higher prediction accuracy and is obviously better than other models.

D. Complexity and Training Time Analysis

A comparison of the number of parameters of the LSTM model, the 3D CNN-LSTM model and the ConvLSTM model, is shown in table 2.

Table 2. Comparison of the number of network parameters and training time for each model

Items	LSTM	3DCNN-LSTM	3DCNN-LSTM-AT	ConvLSTM
Total parameters	16208	69056	277704	3484161
Parameters	16208	69056	277576	3483777
Train Time (Short)	331s	382.5s	637s	965s
Train Time (Mid)	307.5s	526s	1844s	4595s
Train Time (Long)	404.5	552.5s	1343.5s	1922.5s

It can be found that the training parameters of 3D CNN-LSTM-AT model are about 10 times less than those of ConvLSTM, which makes the training much faster and more suitable for prediction task. We also counted the average training time in short-term, medium-term and long-term tasks of each model on the South China Sea and Bohai datasets. As shown in the table above, it can be found that 3DCNN-LSTM-AT has the shorter training time than ConvLSTM. The training time increase slowly with the increase of the prediction step and input data, while the training time is shorter for the long-term prediction due to the cyclic prediction.

V. CONCLUSIONS

SST is a key physical parameter of the world ocean and plays an important role in the air-sea interactions. To achieve an accurate and comprehensive prediction of the SST change over time, a 3DCNN-LSTM-AT Model based on time characteristics is proposed which can capture the correlations of SST across both space and time. A series of experiments on the tasks of the short, medium and long-term SST predictions in Bohai Sea and the South China Sea demonstrate that the proposed model has achieved an ideal effect and outperforms the persistence model. Further improvements to the model will be made later to further improve the prediction accuracy and the training and computational efficiency of the model.

ACKNOWLEDGMENT

This work was supported by the National Key R&D Program of China (NO.2016YFC1401900) and National Natural Science Foundation of China (No. 61872072 and 61073063).

REFERENCES

- [1] Xiao C, Chen N, Hu C, Wang K. A spatiotemporal deep learning model for sea surface temperature field prediction using time-series satellite data. *Environmental Modelling and Software*, 2019, 120.
- [2] H. U. Solanki, D. Bhatpuria, and P. Chauhan. Integrative analysis of AltiKa-SSHa, MODIS-SST, and OCM-chlorophyll signatures for fisheries applications. *Marine Geodesy*, 2015, 38:672-683.
- [3] C. C. Funk and A. Hoell. The leading mode of observed and CMIP5 ENSO-residual sea surface temperatures and associated changes in Indo-Pacific climate. *Journal of Climate*, 28(11):150202132719008.
- [4] Aparna, S.G., D'Souza, S., Arjun. Prediction of daily sea surface temperature using artificial neural networks. *Int. J. Remote Sens.* 2018, 39(12):4214-4231.
- [5] Liu Y, Fu W. Assimilation high-resolution sea surface temperature data improves the ocean forecast potential in the Baltic Sea. *Ocean Science*, 2018, 14:525-541.
- [6] Yang Y, Dong J, Sun X. A CFCC-LSTM Model for Sea Surface Temperature Prediction. *IEEE Geoscience and Remote Sensing Letters*, 2017, 15(2):207-211.
- [7] Yan X, Ants L. Forecasts of tropical Pacific SST and sea level using a Markov model. *Geophys. Res. Lett.* 2000, 27(17):2701-2704.
- [8] I. D. Lins, M. Moura, M. Silva, et al. Sea surface temperature prediction via support vector machines combined with particle swarm optimization. *Proceedings of the 10th International Probabilistic Safety Assessment & Management Conference*, 2010.
- [9] K. Patil, M. C. Deo, and M. Ravichandran. Prediction of sea surface temperature by combining numerical and neural techniques. *Journal of Atmospheric & Oceanic Technology*, 2016, 33(8):1715-1726.
- [10] He Qi, Cha Cheng, Sun Miao, et al. Parallel prediction algorithm for sea surface temperature under Spark platform. *Ocean Bulletin*, 2019, (3):280-289.
- [11] Y. Lecun, Y. Bengio, and G. Hinton. Deep learning, *Nature*, 2015, 521(7553):436-44.
- [12] Zhang Q, Wang H, Dong J. Prediction of Sea Surface Temperature Using Long Short-term Memory. *IEEE Geoscience and Remote Sensing Letters*, 2017, 14(10):1745-1749.
- [13] L. Wei, L. Guan and L. Qu. Prediction of Sea Surface Temperature in the South China Sea by Artificial Neural Networks. *IEEE Geoscience and Remote Sensing Letters*, 17(4):558-562.
- [14] Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., Woo, W.-C. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. *Advances in Neural Information Processing Systems (NIPS 2015)*, 802-810.
- [15] Chen T, He T, Benesty M. Xgboost: Extreme Gradient Boosting. R Package Version 0.4-2, 2015:1-4.
- [16] Peng Y, Tian S, Yu L. A Joint Approach to Detect Malicious URL Based on Attention Mechanism. *International Journal of Computational Intelligence and Applications*, 2019, 18(03):1950021.
- [17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NIPS2017*, 2017, 5998-6008.
- [18] J. Xie, J. Zhang, J. Yu and L. Xu. An Adaptive Scale Sea Surface Temperature Predicting Method Based on Deep Learning with Attention Mechanism. *IEEE Geoscience and Remote Sensing Letters*, 2020, 17(5): 40-744.



Baiyou Qiao received the Ph. D degree in Computer Software and Theory from Northeastern University, China, in 2006. Since 2008, he has been an Associate Professor in Northeastern University, China. He is a member of China Computer Society. His main research interests include cloud computing, Big Data mining and analysis, spatial temporal data management and etc.



Zhongqiang Wu is a postgraduate student of Computer Technology in Northeastern University, China. His main research interests include data mining and time series analysis and prediction.



Zhong Tang is a postgraduate student of artificial intelligence in Northeastern University, China. He received his BS degree from Northeastern University China in June 2020. His main research interests are the ocean Big Data analysis and prediction.



Gang Wu is a Associate Professor in Northeastern University, China. He received BS and MS degrees from Northeastern University China in 2000 and 2003 respectively, and his PhD degree from Tsinghua University in 2008. His main research interests include main memory database, knowledge graph, and social networks. He is a member of ACM, a member of Chinese Information Processing Society of China, and a member of China Computer Federation.