A Hybrid Multivariate Deep Learning Network for Multistep Ahead Sea Level Anomaly Forecasting

GUOSONG WANG, a,b XIDONG WANG, a XINRONG WU, b KEXIU LIU, b YIQUAN QI, a CHUNJIAN SUN, b AND HONGLI FU b

a College of Oceanography, Hohai University, Nanjing, China
b National Marine Data and Information Service, Tianjin, China

(Manuscript received 9 April 2021, in final form 18 November 2021)

ABSTRACT: The accumulated remote sensing data of altimeters and scatterometers have provided new opportunities for ocean state forecasting and have improved our knowledge of ocean-atmosphere exchanges. Studies on multivariate, multistep, spatiotemporal sequence forecasts of sea level anomalies (SLA) for different modalities, however, remain problematic. In this paper, we present a novel hybrid and multivariate deep neural network, named HMnet3, which can be used for SLA forecasting in the South China Sea (SCS). First, a spatiotemporal sequence forecasting network is trained by an improved convolutional long short-term memory (ConvLSTM) network using a channelwise attention mechanism and multivariate data from 1993 to 2015. Then a time series forecasting network is trained by an improved long short-term memory (LSTM) network, which is realized by ensemble empirical mode decomposition (EEMD). Finally, the two networks are combined by a successive correction method to produce SLA forecasts for lead times of up to 15 days, with a special focus on the open sea and coastal regions of the SCS. During the testing period of 2016–18, the performance of HMnet3 with sea surface temperature anomaly (SSTA), wind speed anomaly (SPDA), and SLA data is much better than those of state-of-the-art dynamic and statistical (ConvLSTM, persistence, and climatology) forecast models. Stricter testbeds for trial simulation experiments with real-time datasets are investigated, where the eddy classification metrics of HMnet3 are favorable for all properties, especially for those of small-scale eddies.

KEYWORDS: Sea level; Altimetry; Remote sensing; Operational forecasting; Deep learning

1. Introduction

The South China Sea (SCS) is a large semienclosed sea where typhoons, mesoscale ocean eddies, internal waves, and other weather and marine phenomena occur frequently (Wang et al. 2014; Tuo et al. 2018). The population around the SCS is dense, and hence, the requirements for environmental and operational ocean forecasting are high. In 1992, ERS-1 and T/P remote sensing satellites observed near-global sea level anomaly (SLA) distributions through accurate continuous data, representing the first observing system that could be used for permit eddy-resolution global ocean forecasting (Smedstad et al. 2003). In recent years, various kinds of remote sensing, in situ observation and reanalysis datasets have been produced. However, SLA forecasting has not improved rapidly with the increase in data. SLA forecasting technology can be mainly divided into two types: numerical forecasting and empirical statistical forecasting. For the former, the most widely used global operational ocean models include the Hybrid Coordinate Ocean Model (HYCOM) (Chassignet et al. 2009) and the Nucleus for European Modeling of the Ocean (NEMO) (Madec et al. 2017). Currently, there is a Global National Real-Time Ocean Forecasting System (RTOFS) ocean model, which is based on eddy-resolving 1/12° global HYCOM and provides up to 8 days of forecasts using the daily initialization fields and a 3D multivariate data assimilation approach (Cummings 2006). The French Mercator Ocean International forecast systems use the NEMO to predict sea level values up to 10 days in advance (Drévilleon et al. 2008). With the development of high-performance computing and observation systems, more scientific challenges have surfaced in terms of physical processes, parameterization schemes, and data assimilation algorithms for different numerical models (Bauer et al. 2015).

Deep learning has grown in popularity in recent years and has been applied to subgrid parameterization (Zanna and Bolton 2020), chaotic dynamical system forecasting (Pathak et al. 2018; Vlachas et al. 2018), XBT bias correction (Leahy et al. 2018), and ocean prediction (Zhang and Dai 2019; Song et al. 2020; Berbić et al. 2017). Recent studies have shown that both marine and meteorological forecasts are more accurate and energy efficient in terms of the parameterization of key physical processes (Jiang et al. 2018; Bolton and Zanna 2019; Gentile et al. 2018). Currently, both recurrent neural networks (RNNs) ( Hochreiter and Schmidhuber 1997) and convolutional neural networks (CNNs) (LeCun et al. 1998) have achieved state-of-the-art results on a number of future time series forecasting benchmarks. Ham et al. (2019) used both CMIP5 output and reanalysis with transfer learning to train a CNN, which extended the skillful forecast lead time of the Niño-3.4 index to 1.5 years. Moreover, the CNN model can predict the detailed zonal distribution of sea surface temperatures (SSTs) fairly well.
overcoming a weakness of dynamic forecast models. For forecasting time series, many methods, such as artificial neural networks (Makarynskyy et al. 2004; Imani et al. 2014a), combined least squares–neural network (Zhao et al. 2019), copula-based prediction model (Yavuzdoğan and Tanrı Kayıkçı 2021), and support vector regression (Imani et al. 2014b), are used to forecast the short term of sea level variations. Hybrid models (Fu et al. 2019) can also be a powerful tool to improve the performance by fusing deep features information. Furthermore, an RNN encoder–decoder model based on an attention mechanism also reached a new level (Bahdanau et al. 2015), and the self-attention mechanism has been widely used in recent years (Vaswani et al. 2017; Choromanski et al. 2020).

Deep learning for spatiotemporal sequence forecasting is essential for a wide range of scientific studies and operational applications. Shi et al. (2015) proposed a convolutional long short-term memory (LSTM) (ConvLSTM) structure for spatiotemporal sequences, which converts a step-by-step prediction problem into a spatiotemporal sequence forecast problem with an end-to-end learning framework. Compared with the traditional optical flow method, the accuracy of precipitation forecasting in Hong Kong was significantly improved. The ConvLSTM model has become a seminal work in this area. Subsequently, Ma et al. (2019) constructed an SLA nowcasting network based on ConvLSTM by adopting general learning strategies for iterated multistep (IMS) estimation. IMS estimation learns a one-step ahead forecaster and iteratively applies it to generate multistep predictions to preserve the sharpness of the predicted frames. The results showed that the root-mean-square error (RMSE) of ConvLSTM on the seventh day of the SLA forecast was 3.28 cm, and the matching rate for eddies with diameters greater than 100 km was approximately 60%. Admittedly, the stacked ConvLSTM architecture has been widely used for supervised spatiotemporal learning, where the memory cells that belong to the layers are mutually independent and updated merely in the time domain. Recent advances in ConvRNNs include the introduction of external memory, PredRNN (Wang et al. 2017) and PredRNN++ (Wang et al. 2018), which model spatial and temporal representations in a unified memory cell and convey the memory both vertically across layers and horizontally over states. They have also been shown to be better than ConvLSTM in both radar echo extrapolation and video frame prediction tasks (Wang et al. 2019a).

However, there are three important research challenges for SLA multistep forecasting by multivariate deep learning for different modalities. The first challenge is to improve multistep forecasting for high dimensionality of the spatiotemporal sequences. The IMS approach is easy to train and computationally less expensive, while the direct multistep (DMS) approach, which optimizes multistep forecasting, can avoid the error drifting problem. The selection of DMS or IMS involves a trade-off among forecasting bias, estimation variance, the length of the prediction horizon, and model nonlinearity (Taib and Hyndman 2014). The second challenge is how to propose an effective model for the spatial and temporal structures of different modalities (such as spatial and time series). Recently, hybrid methods, as novel techniques, have become a valuable research topic in both academia and industry (Turki et al. 2015; Alexandre et al. 2015). The aim of combining forecasting methods is to improve the accuracy of predictions by taking advantage of each method. Hybrid methods integrate information from two or more modalities with more comprehensive information characteristics to obtain robust and consistent results (Chen et al. 2019). The third challenge is multivariate spatiotemporal sequence forecasting problems. Most of the models forecast only one channel of images. Regional sea level variability, both temporally and spatially, is dominated by local ocean and atmospheric variability (Miles et al. 2014); thus, SLA forecasting in the SCS is even more challenging due to the chaotic nature of the variability in multiple regional variables. Unfortunately, to the best of our knowledge, state-of-the-art deep learning-based models are seldom applied to such SLA forecasting.

Spatial appearances and temporal variations are two crucial factors for SLA forecasting in the SCS. To tackle the aforementioned challenges, this paper models these structures by presenting a novel hybrid and multivariate deep neural network (HMnet). HMnet, which combines an improved LSTM model and a channel attention-based ConvLSTM model by a successive correction scheme, is proposed for multistep daily SLA forecasting. The proposed HMnet3 method utilizes remotely sensed SSTs, winds, and SLA from the previous 15 days to forecast SLA for the following 15 days, with a special focus on the open sea and coastal regions of the SCS. The core advantages of HMnet are its hybrid structure that extracts and memorizes spatial and temporal representations separately. HMnet is a general framework that can be easily extended to other forecast learning tasks by integrating with other architectures.

2. Data and methods

2.1. Data

In this study, three long-term daily satellite remote sensing products are used for training, testing, and validation in deep learning. The multisatellite altimeter SLA dataset is distributed by the Copernicus Marine Environment Monitoring Service (CMEMS), which provides a consistent and homogeneous catalog of products for near-real-time applications. The gridded wind vector analysis data are from the Cross Calibrated Multi-Platform (CCMP) V2.0 dataset (Atlas et al. 1996, 2011), which provides consistent, gap-free, long-term time series of ocean surface wind vector analysis fields. The SST data are from the NOAA Optimum Interpolation (OI) SST V2 High Resolution Dataset (Reynolds and Chelton 2010). The horizontal resolution of the remote sensing data is 1/4° × 1/4°, and the time period is from 1 January 1993 to 31 December 2018. The SCS domain is (5°–25°N, 105°–125°E).

To obtain the disjoint subsets required for training, testing, and verification, the dataset is divided into three consecutive
sequences according to years: the 21-yr period from January 1993 to December 2013 is used as the training set, the 2-yr period from January 2014 to December 2015 is used as the verification set, and the 3-yr period from January 2016 to December 2018 is used as the test set.

To accelerate the convergence speed of the gradient descent algorithm in HMnet, the satellite-observed sea surface temperature anomaly (SSTA), wind speed anomaly (SPDA), and SLA data in the training set are first normalized to $[-1, 1]$ with minimum–maximum scaling. The land points are set to zero. For the verification and test sets, the original observations are normalized with the same minimum–maximum scaling as the training sets. Then we slice the consecutive observations with a 30-frame-wide sliding window. Thus, each sequence consists of 30 frames, 15 for the input and 15 for forecasting. A total of 9495 sequences were split into a training set of 7669 samples, a validation set of 730 samples and previous time $T$ observations as a series of matrices $\hat{x}_{T+1:T+L}$ contains a set of SLA, SSTA, and SPDA data. In this paper, both the length $L$ and previous time $T$ are 15 days, and the coordinate part of the grid is $80 \times 80$.

\begin{equation}
\hat{x}_{T+1:T+L} = \begin{array}{c}
\arg \max_{x_{T+1:T+L}} \quad p(x_{T+1:T+L} | \alpha_{1:T}) \\
\alpha_{1:T} = \{\text{SLA}_{1:T}, \text{SSTA}_{1:T}, \text{SPDA}_{1:T}\}
\end{array},
\end{equation}

We define the length $L$ of the SLA sequence in the future as a series of matrices $\hat{x}_{T+1:T+L} = [\hat{x}_{T+1:T+2}, \ldots, \hat{x}_{T+L}]$, namely, the observations at previous time $T$ plus the auxiliary observations as a series of matrices $\alpha_{1:T}$; each $\alpha_{1:T}$ contains a set of SLA, SSTA, and SPDA data. In this paper, both the length $L$ and previous time $T$ are 15 days, and the coordinate part of the grid is $80 \times 80$.

b. Methods

Owing to the inherent characteristics of nonlinearity and nonstationarity, the high fluctuations in SLA, and the sharp topographic gradients in the SCS, HMnet integrates ensemble empirical mode decomposition (EEMD), LSTM, and ConvLSTM with attention-based and successive correction methods, and it is proposed to enhance the quality of SLA forecasting. The effectiveness of HMnet is a result of three phases (Fig. 1).

First, an improved channel residual attention-based ConvLSTM network involving DMS (CRAM-ConvLSTM), which is trained by multivariate daily remote sensing observations from $T-14$ to $T$, is used for SLA spatiotemporal sequence forecasting in the SCS. Here, $T$ indicates the forecast initial time. A 3D convolution operation is suitable for spatiotemporal feature learning without the loss of spatial correlation information; it can retain the time information of the input signal and generate the output capacity (Ji et al. 2013). Thus, we use a shallow 3D convolution operation to learn the local short-term spatiotemporal features first. Four CRAM-ConvLSTM layers that apply the DMS approach are stacked to learn the spatiotemporal feature maps of long-term SLA, so an SLA forecast in complex dynamic environments is feasible. The CRAM-ConvLSTM model has $3 \times 3$ convolution kernels and 128, 128, 64, and 32 hidden states. The filter of the first convolution layer is used to detect low-order features, such as eddy edge, angle, and curve features. As the number of convolutional layers increases, the characteristics of eddy motion detected by the corresponding filters become more complicated. We concatenate all the states in the forecasting network and feed them into a $3 \times 3 \times 3$ three-dimensional convolutional layer to generate the final SLA forecast.

FIG. 1. An overview of the proposed HMnet framework. The framework consists of two streams: the SLA time series forecasting network, which is decomposed by the EEMD and trained by an improved LSTM network (EEMD-LSTM), and the SLA spatiotemporal sequence forecasting module with an improved ConvLSTM network and the CRAM. Finally, the two networks are combined by a successive correction method to produce SLA forecasts in the SCS with lead times of up to 15 days.
In addition, we propose an effective channel residual attention module (CRAM) that can be widely applied to ConvLSTM. The structure of CRAM-ConvLSTM is illustrated in Fig. 2. It can be divided into two parts: a CRAM with a 3D CNN and stacked ConvLSTM modules.

This base model is formulated as follows:

\[
\tilde{x}_t = \text{CRAM}(x_t),
\]

\[
i_t = \sigma_m(W_{ii} \ast \tilde{x}_t + W_{hi} \ast H_{t-1} + W_{ci} \ast C_{t-1} + b_i),
\]

\[
f_t = \sigma_m(W_{if} \ast \tilde{x}_t + W_{hf} \ast H_{t-1} + W_{cf} \ast C_{t-1} + b_f),
\]

\[
C_t = f_t \odot C_{t-1} + i_t \odot \sigma(W_{ic} \ast \tilde{x}_t + W_{hc} \ast H_{t-1} + b_c),
\]

\[
o_t = \sigma_m(W_{io} \ast \tilde{x}_t + W_{ho} \ast H_{t-1} + W_{co} \ast C_{t} + b_o),
\]

\[
H_t = o_t \odot \sigma(C_t).
\]

The 4D tensors \(x_t \in \mathbb{R}^{T \times W \times H \times C}, \tilde{x}_t \in \mathbb{R}^{T \times W \times H \times C}, \cdot \ast \in \mathbb{R}^{T \times W \times H \times C}, \cdot \odot \in \mathbb{R}^{T \times W \times H \times C}, \) and \(H_t \in \mathbb{R}^{T \times W \times H \times C},\) which preserve all the temporal and spatial information (time, width, height, and channel information), are inputs, aggregated feature maps, input gates, forget gates, output gates, cell outputs, and hidden states, respectively. \(\mathbb{R} \) denotes the domain of the features. The symbols * and \(\odot \) denote the convolution operation and Hadamard product, respectively, and \(\sigma \) is the sigmoid activation function.

The CRAM can be formulated as follows:

\[
\tilde{x}_t = f_{\text{3D}}[F_t + f_{\text{3D}}(\tilde{x}_t^A + F_t)],
\]

\[
F_t = f_{\text{3D}}(x_t),
\]

\[
\tilde{x}_t^A = f_{\text{pool}}(f_{\text{3D}}(F_t)),
\]

\[
\tilde{x}_t^S = f_{\text{depth}}(f_{\text{3D}}(F_t)),
\]

\[
\tilde{x}_t^F = \sigma(\tilde{x}_t^A + \tilde{x}_t^S) \times f_{\text{3D}}(F_t),
\]

where \(\tilde{x}_t^A, \tilde{x}_t^S, \) and \(\tilde{x}_t^F \) are channel attention, spatial attention, and fused attention, respectively. \(f_{\text{3D}}, f_{\text{pool}}, \) and \(f_{\text{depth}} \) denote the functions of 3D convolution, variance pooling, and depthwise separable convolution for each channel, respectively. Compared with conventional convolution operations, the depthwise separable convolution operation has a smaller parameter size and lower computational consumption.

Second, a novel EEMD-LSTM approach, which combines EEMD and LSTM, is proposed for SLA time series forecasting. EEMD can be performed over all data points. As several previous studies pointed out, scale-separated regression is often superior to all-in-one regression (Wu and Shen 2016).

Step 1: Determine the relative maxima for the RMSE array produced by the CRAM-ConvLSTM method, which returns 20 sites of relative maximum values found in five areas (see Table 1) to construct the SLA time series forecasting networks with the EEMD-LSTM method.
Step 2: The original time series of input $a_t$ from each site, which contain a set of SLA, SSTA, and SPDA time series data, are decomposed into five intrinsic mode functions (IMFs) to obtain more realistic and physically meaningful signals (Huang et al. 2019; Liu et al. 2019); hence, a relatively stationary IMF that can be readily modeled by LSTM is obtained.

Step 3: We use improved multivariate time series forecasting with an LSTM network to fit each IMF from $T_1$ to $T_T$. Then each IMF is forecasted for the SLA time series of $\hat{x}_t$: $\hat{x}_t$ using the corresponding LSTM module.

Step 4: The SLA time series forecasting results are calculated by the sum of the forecasting values of every IMF.

Moreover, the EEMD-LSTM in this paper contains a four-layer LSTM network, which has been used to solve many real-life sequence modeling problems (Sagheer and Kotb 2019; Chao et al. 2018). The time complexity of EEMD-LSTM with hybrid and multivariate networks is larger than that of traditional RNNs. We selected only 20 time series for EEMD-LSTM implementation; as the number of locations increases, the accuracy of HMnet does not improve significantly.

Finally, a successive correction method that integrates multivariate time series forecasting with EEMD-LSTM and multivariate spatiotemporal sequence forecasting with CRAM-ConvLSTM is proposed to improve the quality of SLA forecasting. The successive correction method, which uses the relative scales of the horizontal distance and the SLA difference between the CRAM-ConvLSTM forecasting value and EEMD-LSTM forecasting value, can be expressed as

\[
\phi_{\text{new}}^i = \phi_{\text{old}}^i + \frac{\sum_{k=1}^{N} W_{i,n}^k \left( \phi_{\text{site}}^i - \phi_{\text{new}}^i \right)}{\sum_{k=1}^{N} W_{i,n}^k},
\]

where the subscript $i$ is the index of the grid point and the superscript site represents a single point of the SLA time series, $\phi_{\text{new}}$ represents the revised grid-based forecast, $\phi_{\text{old}}$ represents the grid-based forecast before revision, $\phi_{\text{site}}$ represents the time series forecast, $\phi$ represents the value of $\phi_{\text{old}}$ interpolated to $\phi_{\text{site}}$, $k$ represents the number of time series forecasts within the search range, $i$ represents the index of the spatial grid point, and $N$ represents the total number of time series forecasts. The term $W_{i,n}^k$ represents the mapping operator of the $k$th single point over the $i$th grid point within the search range and is based on the correlation scales and the threshold of the SLA gradient constraint. We can express $W_{i,n}^k$ as

\[
W_{i,n}^k = \exp \left( - \frac{x_i - x_{\text{site}}^n}{L_x} \right)^2 - \frac{|y_i - y_{\text{site}}^n|}{L_y} - \frac{|f_i - f_{\text{site}}^n|}{L_f} \right) \right),
\]

where $x_i$, $y_i$, and $f_i$ are the spatial coordinates of the $i$th grid point; $x_{\text{site}}^n$, $y_{\text{site}}^n$, and $f_{\text{site}}^n$ are the coordinates of the $i$th single point; $L_x$ and $L_y$ represent the correlation scales of the zonal and meridional directions, respectively, and are equal to the lengths of the corresponding directions of this level grid; $f$ stands for the SLA; and $L_f$ is the threshold value of the SLA.

### Table 1. Five sensitive areas, 20 selected sites, and the RMSE values of the two methods used.

<table>
<thead>
<tr>
<th>Sensitive areas</th>
<th>Abbreviation</th>
<th>Lon ('E)</th>
<th>Lat ('N)</th>
<th>ConvLSTM-DMS (cm)</th>
<th>EEMD-LSTM (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>West coast of Taiwan Strait</td>
<td>S1</td>
<td>119.50</td>
<td>24.00</td>
<td>7.95</td>
<td>2.17</td>
</tr>
<tr>
<td></td>
<td>S2</td>
<td>120.00</td>
<td>24.50</td>
<td>7.72</td>
<td>2.18</td>
</tr>
<tr>
<td></td>
<td>S3</td>
<td>118.38</td>
<td>23.88</td>
<td>10.94</td>
<td>2.84</td>
</tr>
<tr>
<td></td>
<td>S4</td>
<td>118.88</td>
<td>24.62</td>
<td>9.76</td>
<td>3.47</td>
</tr>
<tr>
<td>Coast of Guangdong Province</td>
<td>S5</td>
<td>116.50</td>
<td>22.84</td>
<td>6.90</td>
<td>1.44</td>
</tr>
<tr>
<td></td>
<td>S6</td>
<td>114.50</td>
<td>22.00</td>
<td>5.41</td>
<td>1.42</td>
</tr>
<tr>
<td></td>
<td>S7</td>
<td>111.80</td>
<td>21.00</td>
<td>6.27</td>
<td>1.44</td>
</tr>
<tr>
<td></td>
<td>S8</td>
<td>112.50</td>
<td>21.50</td>
<td>7.44</td>
<td>2.40</td>
</tr>
<tr>
<td>Beibu Gulf</td>
<td>S9</td>
<td>108.12</td>
<td>20.38</td>
<td>7.28</td>
<td>2.02</td>
</tr>
<tr>
<td></td>
<td>S10</td>
<td>108.75</td>
<td>21.50</td>
<td>7.42</td>
<td>2.17</td>
</tr>
<tr>
<td></td>
<td>S11</td>
<td>107.12</td>
<td>20.38</td>
<td>6.60</td>
<td>1.80</td>
</tr>
<tr>
<td></td>
<td>S12</td>
<td>107.20</td>
<td>19.50</td>
<td>6.00</td>
<td>1.50</td>
</tr>
<tr>
<td></td>
<td>S13</td>
<td>107.50</td>
<td>18.50</td>
<td>5.48</td>
<td>1.15</td>
</tr>
<tr>
<td></td>
<td>S14</td>
<td>108.20</td>
<td>18.00</td>
<td>5.32</td>
<td>1.10</td>
</tr>
<tr>
<td></td>
<td>S15</td>
<td>109.50</td>
<td>20.50</td>
<td>8.15</td>
<td>2.16</td>
</tr>
<tr>
<td>East of Luzon Strait</td>
<td>S16</td>
<td>124.62</td>
<td>19.62</td>
<td>5.02</td>
<td>1.05</td>
</tr>
<tr>
<td></td>
<td>S17</td>
<td>124.62</td>
<td>21.50</td>
<td>5.10</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>S18</td>
<td>124.88</td>
<td>22.12</td>
<td>5.97</td>
<td>0.95</td>
</tr>
<tr>
<td>West of Luzon Strait</td>
<td>S19</td>
<td>119.85</td>
<td>21.85</td>
<td>5.07</td>
<td>1.05</td>
</tr>
<tr>
<td></td>
<td>S20</td>
<td>118.63</td>
<td>21.12</td>
<td>9.59</td>
<td>1.64</td>
</tr>
</tbody>
</table>
difference between the single point and the grid point and is set to 0.1 m in this study.

The correction process is independent of features, and errors from different networks are usually irrelevant and will not cause error accumulation. The late fusion scheme tends to have better performance for most concepts (Snoek and Worring 2005). On the basis of deep learning, we use a successive correction method based on a strong gradient constraint to perform rapid late fusion. The successive correction scheme effectively prevents excessive unphysical projections and thus greatly improves the quality of the forecast (Fu et al. 2013).

We train all deep learning networks by minimizing the mean squared error loss of time back propagation, and the optimizer is Adam with a learning rate of 0.0001 (Zeiler 2012).

L2 regularization makes the network use all input features to prevent the neural network from being overfitted. The small batch gradient descent method is used to update the parameters and to reduce randomness and computational costs. The batch size of each iteration is set to 32, and the parameters are separately adjusted to obtain good performance. All experiments are implemented in TensorFlow on an NVIDIA Quadro P6000 GPUs with 50 G of memory. We perform early stopping and directly optimize the 15-day ahead daily SLA forecast.

3. Results and evaluation

In this paper, we use the RMSE and the Pearson correlation coefficient (PCC) to evaluate the performance of the network:

\[
\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (X_{\text{forecast},i} - X_{\text{obs},i})^2},
\]

\[
\text{PCC} = \frac{\sum_{i=1}^{N} (X_{\text{forecast},i} - \bar{X}_{\text{forecast}})(X_{\text{obs},i} - \bar{X}_{\text{obs}})}{\sqrt{\sum_{i=1}^{N} (X_{\text{forecast},i} - \bar{X}_{\text{forecast}})^2 \sum_{i=1}^{N} (X_{\text{obs},i} - \bar{X}_{\text{obs}})^2}},
\]

where \(N\) is the total number of test sets, \(X_{\text{forecast}}\) is the forecasted SLA, and \(X_{\text{obs}}\) is the satellite altimeter SLA dataset. \(\bar{X}_{\text{forecast}}\) is the average forecasted SLA, and \(\bar{X}_{\text{obs}}\) is the average satellite altimeter SLA. The PCC values are between \(-1\) and 1.

a. Assessment of the hybrid SLA forecasting model

First, we compare the performances of CRAM-ConvLSTM and EEMD-LSTM using the test sets. We also include widely used univariate ConvLSTM networks, which follow the IMS and DMS structures, as the baseline models (ConvLSTM-IMS and ConvLSTM-DMS).

Figure 3 shows the mean area-weighted RMSE of the SLA during different lead days produced by the three methods using the 2016–2018 test sets. We can see that both the ConvLSTM-DMS and ConvLSTM-IMS networks can learn the temporal and spatial relationships in the central SCS. For the first 7 lead days, the large RMSE values are mainly located in the coastal regions of the northern SCS, which may be attributed to the relatively large dynamic height caused by coastal currents and Ekman transport. For the
last 8 lead days, the RMSE in the Luzon Strait, where mesoscale eddies are active (Wang et al. 2012), grows rapidly. Compared to the ConvLSTM-IMS network, the ConvLSTM-DMS network can reduce the SCS averaged RMSE during all lead days by 17.3%, i.e., from 6.17 to 5.10 cm.

The spatial structure and magnitude of the RMSE produced by ConvLSTM-DMS are similar to those produced by ConvLSTM-IMS. Without time series analysis, it is difficult to extract features in different regions for ConvLSTM. A possible reason is that the basic ConvLSTM architecture uses a fixed kernel, which can only capture a single-modal state (Shi et al. 2015). It is almost impossible for a single-modal deep learning network, such as ConvLSTM, to make accurate and comprehensive SLA forecasts at the same time.

The SCS has a wide continental shelf, a deep sea basin, and steep continental slopes. Such topographic changes have considerable impacts on the forecasting ability of the network (Masina and Pinardi 1994). Thus, it is vital to validate the model results in both coastal regions and open-ocean regions. Figure 4 shows strong nonlinearity and nonstationarity in the shelf shallower than 100 m and in the Luzon Strait, which has a strong impact on the performance of the network. Over the shelf area in the SCS, however, the SLA data still contain aliases from tides and internal waves (Yuan et al. 2006). Since the error distribution of SLA is in line with the area where the sea level changes dramatically in the SCS, it is necessary to develop an in situ forecasting network to decrease the RMSE in the shelf area and Luzon Strait. Therefore, we determined the coastal regions of the northern SCS (the west coast of the Taiwan Strait, the coast of Guangdong Province, and Beibu Gulf) and the Luzon Strait (east of the Luzon Strait and west of the Luzon Strait) to be the sensitive areas. A detailed description of sensitive areas is given in appendix A.

After CRAM-ConvLSTM, extrema of the RMSE field in the sensitive areas are detected to estimate the local

---

**FIG. 4.** An example of 15-day SLA forecasting of ConvLSTM-DMS and EEMD-LSTM against the remote sensing test sets for 20 selected sensitive sites.
maximum SLA forecasting locations. The average RMSE on the test sets of each time series forecast shows that the EEMD-LSTM network reduces the RMSE of the CRAM-ConvLSTM network by 74.9%, i.e., from 6.97 to 1.75 cm. Significant improvements were obtained due to the reduction in mismatches compared with CRAM-ConvLSTM. Some SLA comparisons (Fig. 2) clearly illustrate these improvements.

Since each deep learning network is sensitive to some conditions, another advantage of combining the networks is to reduce bias. HMnet, which integrates the forecasting results of EEMD-LSTM at the 20 sites into those of the CRAM-ConvLSTM network using the successive correction method, is proposed to enhance the quality of SLA forecasting. The statistics are also substantially improved by using HMnet to lower the errors in the coastal region of the northern SCS and Luzon Strait over the full 15-day forecast (Figs. 3m–r). This suggests that no single-modal deep learning network contains all of the sea level information, and the hybrid forecasting model is simple and quite efficient in gaining the shared representative information among different networks to resolve the problem of SLA forecasting.

b. Sensitivity analysis of multivariate forecasting

Altimetry SLA, combined with multiple atmospheric and oceanic variables, wind and SST, can help SLA forecasting. However, the application of multiple input variables for SLA forecasting remains a key challenge. Here, we validate the effectiveness of the CRAM through multivariate sensitivity tests. Two satellite remote sensing datasets, SST and SPDA, are included as additional input variables in the tests. We train three four-layer HMnet models with 1 (SLA), 2 (SLA and SST), and 3 (SLA, SST, and SPDA) input variables. It is worth noting that the networks use the same model structure and the best parameters; therefore, we can impartially compare the different experiments.

As seen from the RMSE and PCC of the sensitivity experiments on 2016–18 data (Table 2), HMnet3 is significantly superior to the ConvLSTM-IMS network. HMnet2 reduces the RMSE produced by HMnet1 from 2.918 to 2.794 cm. HMnet3 further reduces the RMSE produced by HMnet2 from 2.794 to 2.784 cm. As the lead time increases, the PCC gradually decreases. The PCC of HMnet3 gradually decreases from 0.879 for the 1-day lead time to 0.831 for the 15-day lead time. The median PCC is 0.86, and the median RMSE is 0.89. HMnet3 is slightly better than HMnet2. We find that HMnet with the multivariate (HMnet2 and HMnet3) method produces a smaller RMSE than HMnet1 for all years. Although multivariate analysis may not be able to resolve the physical process, the more input variables there are, the greater the improvement in the SLA forecasting skills is.

Figure 5 presents the corresponding step-by-step quantitative comparison to evaluate the forecasting results. We can see that the HMnet models perform much better than the ConvLSTM-IMS model for almost all start times and lead days. There are no significant differences among the three HMnet models, which might be because they have the same convolution kernel size or hidden state layer number. However, compared to HMnet1, the majority of the improvement is seen in October. In particular, the RMSE of the day 12 (day 13) forecasts of HMnet2 (HMnet3) is comparable to the RMSE of the day 10 forecasts of HMnet1.

To investigate the spatial forecasting skills of the HMnet models, Fig. 6 illustrates the spatial variation. There are no obvious differences in the first 4 lead days among the three HMnet models. For lead times longer than 7 days, the HMnet2 and HMnet3 models gradually reduce forecasting errors in the central SCS and East Luzon Strait. Without marine and meteorological forcing, HMnet1 can only obtain a rough and smooth SLA pattern. In contrast, by introducing the CRAM, which enhances the perception of interaction between variables, both HMnet2 and HMnet3 divide typical atmospheric and ocean factors into different feature layers, producing relatively stable SLA forecasts. Note that HMnet3 has a performance that is similar to that of HMnet2 for the first 12 lead days. Compared to HMnet2, HMnet3 has advantages in the central SCS, East Luzon Strait, and Taiwan Strait for longer lead times.

c. Comparisons with various models

A stricter test bed, the advanced operational deep learning forecasting system using HMnet3, is developed. Using the same test sets from August 2019 to August 2020, we compare it with the RTOFS model, which assimilates altimetry SLA observations, and three sets of benchmark forecast models, including two statistical forecast models (persistence and daily climatology) and ConvLSTM-IMS.

The persistence model assumes that the forecast initial state persists for the entire lead time (Levine and Wilks 2000), and it represents an economic forecasting system (Shriver et al.)
The daily forecasted SLA from the RTOFS is compared with remote sensing altimeter data. The mean sea surface (MSS) used in RTOFS is the 1993–2012 temporally averaged sea surface height (SSH) referred to as the Geoid. The MSS is interpolated to a grid of 0.25° × 0.25°.

By observing 15 days of forecasting examples of different models (Fig. 7), we can see that HMnet3 not only significantly reduces the error on the coast of the SCS but also forecasts the propagation and evolution of the anticyclonic eddy shed in the Luzon Strait. This is mainly due to two reasons. One is that HMnet3 can fit open boundary conditions. There are a large number of eddy motion samples in the training sets in the Luzon Strait. HMnet3 can learn the eddy spatiotemporal characteristics of different regions with different networks during training and remember detailed appearances and long-term motions; thus, it can make reasonable forecasts at the boundary. The other is that HMnet3 is forced by marine and atmospheric fields using the CRAM for end-to-end training, and the eddy structure and intensity forecast can be improved, especially for the last 7 lead days.

ConvLSTM-IMS is less computationally expensive in training, but the accumulation of errors is caused by using the forecasting result of the previous time step as the input of the next time step in multiple iterations. Therefore, the initial error in the southern Beibu Gulf in this example has a great effect on the errors in the day 8 to day 15 forecasts (Figs. 7O–R). Moreover, the complex structure of ConvLSTM-IMS is still plagued by the problem that the gradient of the cost function disappears during training. Through the back propagation of time, the amplitude of the gradient decays exponentially. The dependence on long-term prediction and training can easily cause the problem of the decay in the SLA signal. As a result, the forecasting probability of identifying irregularly shaped features as eddies in the central and western SCS decreases for time frames that are more than one week. Although the RTOFS and persistent forecasting models alleviate the
forecasted shape error compared to the deep learning networks, they produce more false forecasts and less accurate results than the deep learning networks, especially in the shelf areas that are shallower than 100 m and the Luzon Strait, where the sea level changes dramatically.

Figure 8 presents the corresponding step-by-step quantitative comparison of the above five models on the 1-yr trial run. Climatology has poor forecasting performance, while persistence forecasting and ConvLSTM-IMS have certain forecasting skills up to 7 days ahead. Consistent with previous results (Xu et al. 2011), the 8-day forecast mean of the area-weighted RMSE of the HYCOM reached 14 cm across the whole SCS, mainly because the errors in the MSS have significant impacts on the forecasting results. Since the HYCOM does not assimilate the SSH anomaly field directly, it is diagnosed from the prognostic bottom pressure and internal density fields (Halliwell et al. 2014). The MSS must match that contained in the time-mean altimeter data, which is a nontrivial problem (Cummings and Smedstad 2014).

Overall, HMnet3 outperforms all the baseline models and shows superior forecasting skills. This is mainly because HMnet3 is trained end to end for SLA forecasting, and the spatiotemporal convolution structure of the network successfully learns the spatiotemporal sequence forecast characteristics of SLA. For statistical forecasting methods, it is difficult to find a reasonable way to update the future SLA field and train all data end to end.

4. Discussion

a. Estimation of eddy characterization and tracking

Here, we investigate the eddy forecasting performance of the directly forecasted SLA field. After removing larger-scale variability using a low-pass filter, extrema of the SLA field are detected to estimate eddy locations and properties. We apply a new SSH-based eddy tracking tool to identify and automatically track oceanic eddies (Zeiler 2012) and to calculate their radii and amplitudes. By tracing the specific trajectories of several mesoscale eddies, the forecasting performance of the deep learning networks for the movement and propagation of mesoscale eddies in the SCS is discussed.

We select the eddies on 9 May 2016, which include three anticyclonic eddies (AE1–AE3) and three cyclonic eddies (CE1–CE3), as an example. A qualitative comparison of the trajectories is given in Fig. 9, and quantitative comparisons of the amplitudes and locations are given in Fig. 10.

Satellite observations show that AE1 and CE1 in the eastern part of the Luzon Strait moved northward under the influence of the Kuroshio. They moved 214 and 122 km, respectively. AE1 was a warm eddy with a small amplitude and radius compared to the other eddies and disappeared on the fifteenth lead day. ConvLSTM-IMF forecasted that AE1 would move northward, its amplitude and location errors would gradually increase, and it would finally disappear on the eleventh lead day. ConvLSTM-IMF forecasted that CE1 would move westward, while the satellite-observed CE1 moved northward. The amplitude and location errors of CE1 forecasted by ConvLSTM-IMF are 8 cm and 52.21 km, respectively, which are much larger than those forecasted by HMnet3 (i.e., 4.10 cm and 24.73 km).

For CE2, the ConvLSTM-IMF and HMnet3 forecasted eddies disappear on the ninth and twelfth lead days, respectively. The amplitude (location) errors are 4.67 cm (6.24 km) and 3.75 cm (9.90 km) for ConvLSTM-IMF and HMnet3, respectively. For CE3, the ConvLSTM-IMF forecasted eddy moved northwestward, which deviates from the satellite observations. In contrast, the moving direction of CE3 forecasted by HMnet3 is consistent with the observation, and the 15-day averaged errors of the amplitude and position are only 1.38 cm and 14.08 km, respectively. The results of the two cyclonic eddies show that HMnet3 has a certain forecasting ability for small-scale eddies and is better than ConvLSTM-IMF.

AE2 and AE3 originated from a warm eddy generated in the midwestern area of the SCS on 15 March. The original eddy moved northward and broke off, forming AE2 and AE3 on 28 April. Then AE3 moved toward the southwest,
while AE2 moved toward the northwest. AE3 had an irregular oval shape with a long-axis diameter of more than 700 km. From 9 to 23 May, AE2 and AE3 merged into an anticyclonic eddy, which contained two warm cores, and then further formed a nearly circular anticyclonic eddy. At this time, the intensity of AE3 reached its maximum. The period shown in this paper corresponds to the separation and merging processes. In this case, it is difficult for the network to forecast the eddies accurately. According to Figs. 9 and 10, AE2 has no obvious movement in the forecast. The amplitude and location errors of ConvLSTM-IMS (HMnet3) are 3.60 (2.10) cm and 14.20 (12.80) km, respectively. HMnet3 forecasted that AE3 would disappear on the thirteenth lead day, while ConvLSTM-IMS failed on the fourth lead day, and the eleventh to fifteenth lead day forecasting results had large location errors (>100 km). It is worth noting that the shape error of the ConvLSTM-IMS forecast in the central and western SCS increases for more than one week, as illustrated by the blurry forecast in Figs. 7M–R, so many false eddies are generated. Therefore, the

**FIG. 7.** An example of 15-day SLA forecasting (cm) of different models on 6 Dec 2019 using (A)–(F) altimeter observation, (G)–(I) HMnet3, (M)–(R) ConvLSTM-IMS, (S)–(X) Persistence, (a)–(f) Climatology, and (g)–(i) RTOFS.
identified center locations, amplitudes, and radii of AE3 and AE2 are misleading.

b. Eddy forecasting evaluation

To further verify the experimental results, the forecast results of the eddies are evaluated, and 365 trial run experiments from August 2019 to August 2020 spanning from the first to the fifteenth day are used to quantify eddy properties. A series of metrics are used to demonstrate the quality of the forecasted eddies by various models made by the operational deep learning forecasting system using HMnet3. These scores are associated with the identification of the eddy structure and trajectory from those of noneddies. The classification metrics are defined as follows:

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},
\]

\[
\text{FAR} = \frac{FP}{TP + FP},
\]

\[
\text{TS} = \frac{TP}{TP + FN + FP},
\]

\[
\text{Recall} = \frac{TP}{TP + FN},
\]

where FP and FN represent the number of false forecasts of eddies and noneddies, respectively. TP and TN represent the number of true forecasts of eddies and noneddies, respectively.

Intuitively, the accuracy and threat score (TS) can be interpreted as the proportion of correct forecasts. TN is zero in this classification task, so the accuracy and TS are the same. The false alarm rate (FAR) is the ability of the networks to incorrectly forecast a noneddy as an eddy, and recall is the ability to correctly forecast all of the eddies.

The radii of these eddies observed by satellite in the SCS range from approximately 34.9 to 226.0 km, with a mean value of 95.4 km. Eddies with a radius greater than 100 km account for 42.4% of the eddies, and the occurrence probabilities of cyclonic and anticyclonic eddies are approximately the same (Xiu et al. 2010; Chen et al. 2011). The average amplitudes are 213.1 cm for cyclonic eddies and 8.7 cm for anticyclonic eddies; the radius of the cyclonic eddies (98.5 km) is slightly larger than that of the anticyclonic eddies (92.1 km).

Figure 11 shows the statistical characteristics of the classification metrics and lead times on the forecasting performance. The eddy forecasts of various models are compared against altimeter observations by using the same eddy identification method. Although their eddy identification rate is highly accurate, persistence forecasts are not suitable for eddy forecasting due to their high FARs and relative differences in amplitude predictions. Similarly, ConvLSTM-IMS is not suitable for eddy identification due to prediction deviations in the center locations. The HMnet3 forecasts are generally very close to the altimetry observations, and HMnet3 eddy forecasts are significant for all properties. The rates of cyclonic eddies and anticyclonic eddies identified by HMnet3 are 98.4% and 97.3%, respectively, on the first lead day compared to the altimetry observations and 49.0% and 82.9%, respectively, on the fifteenth lead day.

Furthermore, the number, amplitude, and duration of the mesoscale eddies identified by the RTOFS are smaller than those identified by satellite altimetry observations. The
average recall rates of cyclonic eddies (anticyclonic eddies) with radii less than 100 km by RTOFS and HMNet3 are 14.5% (9.6%) and 42.6% (29.5%) on day 8, respectively. When the radii increased to more than 100 km, the recall rates increased to 25.1% (17.9%) and 55.5% (52.3%), respectively. The relative differences in amplitude and center location generated by the RTOFS are also considerably larger than those generated by HMNet3. For eddies with weak kinetic energy and small radii, the RTOFS struggles to accurately reproduce and predict their center locations. However, HMNet3 has a better prediction ability for both small- and large-radius eddies. There are noticeable differences between HMNet3 and RTOFS in quantifying eddy properties, highlighting the potential of purely satellite data-driven deep learning models for oceanic eddy forecasting (Zheng et al. 2020).

c. Limitations

One potential disadvantage of HMNet3 may be the high computational complexity cost of matrix operations for each CRAM-ConvLSTM layer. The parameters of the CRAM-ConvLSTM were optimized over 200 epochs to minimize the loss on the training dataset, which corresponds to approximately 12,000 min of computing time. Therefore, the domain size used in the experiment in this paper is relatively small, so it is acceptable for attentional computation. This could be a limiting factor for a large range of domain. We only generated SLA forecasts up to 15 days and selected 20 time series for HMNet implement. Given the global application needs, further work is needed to determine the impact of longer lead times and more time series. Finally, technical implementation details of the hyperparameter search are given in appendix B.

5. Conclusions

In this paper, we presented a novel HMNet model for short-term SLA forecasting with high accuracy in the SCS. The results showed that HMNet3, which integrates two deep learning networks (CRAM-ConvLSTM and EEMD-LSTM) by successive correction methods, could effectively produce 15-day ahead daily SLA forecasts within seconds. Benefiting from the hybrid improvements in multivariate participates, HMNet3 could greatly reduce the errors in both coastal regions and the Luzon Strait with its long, stable, and accurate spatiotemporal sequence forecast. Compared to the ConvLSTM-IMS network, HMNet3 can reduce the SCS averaged RMSE during all lead days by 25.1%. With these forecasted SLA fields, mesoscale eddies were further identified, and HMNet3 showed skillful capabilities with higher accuracy than current state-of-the-art dynamic or data-driven approaches. The average recall rate of eddies with radii less than 100 km by HMNet3 (36.09%) was more than 3 times higher than that by RTOFS (12.05%) on day 8. HMNet3 was capable of forecasting the propagation and evolution of eddies, and the continuous eddy trajectories forecasted by HMNet3 indicate better resolving capabilities, especially for small-scale eddies.

Our study is a step toward using deep learning networks to extend the reach of oceanic eddy forecasting. Our proof-of-concept study was conducted in a purely satellite data-driven manner. HMNet3 for multistep ahead SLA forecasting is skillful and has high potential in solving real-world problems such as operational marine forecasting and disaster protection problems. This new forecasting technique can potentially be concurrently applied to ocean research and

FIG. 10. Comparison of the amplitudes and radii of HMNet3 (blue) and ConvLSTM-IMS (green) on 9–24 May 2016. The solid lines represent the amplitude error (left axes), and the dotted line represents the location error (right axes).
fisheries. Thus, based on improved synthetic ocean profiles (ISOPs) (Townsend et al. 2015) and deep learning, such as CNNs, future research may examine the relationship between sea surface information and the state of the subsurface ocean. The combination of physical theory and machine learning could prove more effective than either component in isolation.

Acknowledgments. This work was supported by the National Key Research and Development Program of China, Grant 2019YFC1510000; the National Natural Science Foundation of China, Grants 41976019, 41776004, and 41906009; and National Programme on Global Change and Air-Sea Interaction of China. We thank the Copernicus Marine Service for the altimeter satellite gridded SLA data in near–real time (https://marine.copernicus.eu), the Remote Sensing Systems for near-real-time version 2.0 data from the CCMP Ocean Surface Wind Velocity Data set (www.remss.com), and the NOAA/OAR/ESRL PSL, Boulder, Colorado, for the NOAA High Resolution SST data (https://psl.noaa.gov/). The altimetric Mesoscale Eddy Trajectory Atlas products were produced by SSALTO/DUACS and distributed by AVISO+ (https://www.aviso.altimetry.fr) with support from CNES, in collaboration with Oregon State University with support from NASA.

APPENDIX A

Sensitive Areas

Generally, sea level changes dramatically in the area where the water depth is shallower, and the sea area northwest of Luzon Strait in the SCS. Both SLA numerical forecasts and deep learning forecasts result has shown (Ma et al. 2019) the extreme errors are mainly distributed in line with the area where the sea level changes dramatically in the SCS region. In this paper, for the first 7 lead days, the large RMSE values are mainly located in the coastal regions of the northern SCS (Fig. 3), which may be attributed to the relatively large dynamic height caused by coastal currents and Ekman transport. For the last 8 lead days, the RMSE in the Luzon Strait, where mesoscale eddies are active (Wang et al. 2012), grows rapidly (Fig. 3). Therefore, we determined the coastal regions of the northern SCS (the west coast of the Taiwan Strait, the coast of Guangdong Province, and Beibu Gulf) and the Luzon Strait (east of the

Fig. 11. Statistical characteristics of the (top left) relative difference in amplitudes, (top right) eddy identification rates, (middle left) FAR, (middle right) recall, and (bottom) accuracy (TS) of eddy forecasting using various models against altimeter observations over the SCS during the period of August 2019–August 2020. The black curve represents the HMnet3 forecast, the red curve represents the ConvLSTM-IMS forecast, the blue curve represents the persistence forecast, and the green curve represents the RTOFS forecast.
Luzon Strait and west of the Luzon Strait) to be the sensitive areas (Fig. A1).

APPENDIX B

Hyperparameter Search

We mainly optimize the two hyperparameters of layers and kernels size, and the learning rate is set $1 \times 10^{-2}$. The minibatch size is set to 32.

As stacking two to four recurrent layers is a common practice (Wang et al. 2019b), we test four variants of CRAM-ConvLSTM model with different number of layers with different hidden states (Table B1). All the input-to-state and state-to-state kernels are of size $3 \times 3$. Our experiments show that the four-layer CRAM-ConvLSTM networks perform better than the other CRAM-ConvLSTM networks with different numbers of layers, although the improvement is not significant. As the number of layers increases, parameters increases significantly, leading to more training difficulty and more training time.

On the other hand, the size of the kernel is important to capture the spatiotemporal correlations of SLA and eddy. Therefore, we performed another comparative experiment on the performance impact of different kernel sizes ($1 \times 1$, $3 \times 3$, and $5 \times 5$) on three- and four-layer CRAM-networks. The comparison results in Table B1 show that the results become much worse with the $1 \times 1$ kernels size, because it is difficult to capture the spatiotemporal motion patterns (Shi et al. 2015). Meanwhile, the difference between the results of the $3 \times 3$ and $5 \times 5$ kernels size is not significant, but the $3 \times 3$ kernels size has 34% fewer parameters than the $5 \times 5$ kernels size.

We compare EEMD-LSTM with stacked-LSTM modules that have two LSTM layers, three LSTM layers, four LSTM layers, and five LSTM layers. The performance of stacked-LSTM modules is shown in Table B2. As shown in Table B2, the forecast accuracy increases with the number of layers from two to four, and then decreases when the number of LSTM layers increases from four to five. Few layers cannot mine the rules of SLA data well due to inadequate nonstationary modeling capability, while too many layers eventually cause a worse effect due to the vanishing gradient problem and training difficulties.

REFERENCES


Bahdanau, D., K. Cho, and Y. Bengio, 2015: Neural machine translation by jointly learning to align and translate. Third
Int. Conf. on Learning Representations, San Diego, CA, ICLR.


